

# Geomatics for Environmental Management: An Open Textbook for Students and Practitioners



# Contents

<b>About this textbook</b>	<b>15</b>
Who Should use this textbook . . . . .	16
How to adopt this textbook in your geomatics classroom . . . . .	16
How to get involved . . . . .	17
Acknowledgements . . . . .	17
<b>1 What is Geomatics?</b>	<b>19</b>
Learning Objectives . . . . .	19
Key Terms . . . . .	19
1.1 The Science and Technology of Geomatics . . . . .	20
1.2 Information Systems . . . . .	20
1.3 The Five Components of GIS . . . . .	21
1.4 What a GIS can do . . . . .	23
1.5 Modelling the world with GIS . . . . .	24
1.6 A Canadian History of GIS . . . . .	25
1.7 Summary . . . . .	36
Reflection Questions . . . . .	36
Practice Questions . . . . .	36
<b>2 Mapping Data</b>	<b>39</b>
Learning Objectives . . . . .	39
Key Terms . . . . .	39
2.1 Introduction to Geodesy . . . . .	40
2.2 Models of Earth . . . . .	41
2.3 Case Study: The Canadian Geodetic Vertical Datum of 2013 . .	48
2.4 Referencing Location . . . . .	48
2.5 Map Projections for Environmental Management . . . . .	56
2.6 Summary . . . . .	61
Reflection Questions . . . . .	62
Practice Questions . . . . .	62
<b>3 Data Types and Spatial Data Models</b>	<b>63</b>
3.1 Types of Phenomena . . . . .	64
3.2 Types of Data . . . . .	67

3.3	Spatial Is Special . . . . .	77
3.4	Spatial Data Models . . . . .	77
3.5	Choice of Spatial Data Model . . . . .	88
3.6	Case Study . . . . .	89
3.7	Summary . . . . .	89
<b>4</b>	<b>Collecting and Editing Data</b>	<b>91</b>
	Learning Objectives . . . . .	91
	Key Terms . . . . .	91
4.1	Open Data . . . . .	92
4.2	Finding Data . . . . .	92
4.3	Data in Academia . . . . .	93
4.4	Government Data . . . . .	94
	Your Turn! . . . . .	94
4.5	Census Data . . . . .	94
4.6	Census of Canada Geographic Levels . . . . .	95
	Call Out . . . . .	97
4.7	Accessing Census Data . . . . .	97
4.8	Non-Governmental Organization Data . . . . .	98
4.9	Citizen Science . . . . .	98
	Your Turn! . . . . .	99
4.10	International Data . . . . .	99
4.11	Metadata . . . . .	99
4.12	Unpublished Data and the Data Request . . . . .	99
4.13	Historical Data Collections . . . . .	100
4.14	Historical Aerial Photographs . . . . .	100
4.15	Accessing Historical Aerial Photograph Collections . . . . .	102
	Your Turn! . . . . .	102
4.16	Natural Resource Administrative Data . . . . .	104
4.17	Historical Maps . . . . .	105
4.18	Georeferencing Historical Maps . . . . .	105
4.19	Summary . . . . .	108
4.20	Reflection Questions . . . . .	108
4.21	Practice Questions . . . . .	108
4.22	Summary . . . . .	108
<b>5</b>	<b>Relational Databases</b>	<b>109</b>
	Learning Objectives . . . . .	109
	Key Terms . . . . .	109
5.1	Relational Database Management Systems . . . . .	110
5.2	Relational Databases . . . . .	110
5.3	Relational Algebra . . . . .	112
5.4	Selection . . . . .	113
5.5	Projection . . . . .	113
5.6	Rename . . . . .	114
5.7	Set Union . . . . .	114

5.8 Set Intersection . . . . .	115
5.9 Set Difference . . . . .	115
5.10 Cartesian Product . . . . .	116
5.11 Set Division . . . . .	117
5.12 Boolean Algebra . . . . .	120
5.13 Equality Operators . . . . .	120
5.14 Conditional Operators . . . . .	122
5.15 Joining Relations . . . . .	124
5.16 Keys . . . . .	126
5.17 Natural Join . . . . .	127
5.18 Outer Join . . . . .	128
5.19 Right and Left Outer Join . . . . .	128
5.20 Theta Join . . . . .	129
5.21 Cardinality of Joins . . . . .	129
5.22 Structured Query Language . . . . .	129
5.23 Case Study: Combining Socioeconomic and Vegetation Information for Assessing Population Vulnerability . . . . .	133
5.24 Join . . . . .	133
5.25 Calculation . . . . .	134
5.26 Query . . . . .	136
Remember This? . . . . .	136
5.27 Summary . . . . .	137
Reflection Questions . . . . .	137
Practice Questions . . . . .	137
Recommended Readings . . . . .	137
<b>6 Overlay and Proximity Analysis</b> . . . . .	<b>139</b>
Learning Objectives . . . . .	139
Key Terms . . . . .	139
6.1 Cartographic Modelling . . . . .	140
6.2 Overlay Methods . . . . .	144
6.3 Proximity Methods . . . . .	144
6.4 Summary . . . . .	145
Reflection Questions . . . . .	145
Practice Questions . . . . .	145
Recommended Readings . . . . .	145
<b>7 Topology and Geocoding</b> . . . . .	<b>147</b>
Learning Objectives . . . . .	147
Key Terms . . . . .	147
7.1 Topology . . . . .	148
7.2 Planar vs. Non-Planar Topology . . . . .	149
7.3 Implementing Planar Topology . . . . .	150
7.4 Adjacency and Overlap . . . . .	150
7.5 Intersect and Connect . . . . .	151
7.6 Coincident and Disjoint . . . . .	154

7.7	Cover . . . . .	154
7.8	3D topologies . . . . .	163
7.9	Geocoding . . . . .	168
7.10	Case Study: Working with Canadian Census Data . . . . .	170
7.11	Summary . . . . .	170
	Reflection Questions . . . . .	170
	Practice Questions . . . . .	171
	Recommended Readings . . . . .	171
<b>8</b>	<b>Network Analysis</b>	<b>173</b>
	Learning Objectives . . . . .	173
	Key Terms . . . . .	173
8.1	Introduction to Graph Theory . . . . .	174
8.2	Nodes . . . . .	174
8.3	Edges . . . . .	174
8.4	Connectivity and Order . . . . .	174
8.5	Direct . . . . .	174
8.6	Undirect . . . . .	176
8.7	Network Topologies . . . . .	176
8.8	Physical vs. Logical Topology . . . . .	176
8.9	Non-Hierarchical Topologies . . . . .	178
8.10	Hierarchical Topologies . . . . .	180
8.11	Spatial Network Analysis . . . . .	183
8.12	Network Tracing . . . . .	184
8.13	Linear Referencing . . . . .	185
8.14	Routing . . . . .	185
8.15	Least Cost Paths . . . . .	185
8.16	Least Cost Corridors . . . . .	186
8.17	Reach Analysis . . . . .	187
8.18	Network Centrality . . . . .	187
8.19	Closeness Centrality . . . . .	188
8.20	Betweenness Centrality . . . . .	188
8.21	Case Study: Central and Peripheral Green Spaces in Vancouver .	191
	Practice & Reflection . . . . .	192
<b>9</b>	<b>Raster Analysis and Terrain Modelling</b>	<b>193</b>
	Learning Objectives . . . . .	193
	Key Terms . . . . .	193
9.1	Raster Analysis . . . . .	194
9.2	Digital Vertical Models . . . . .	194
9.3	Digital Elevation Models (DEM) . . . . .	195
9.4	Digital Terrain Models (DTM) . . . . .	196
9.5	Digital Surface Models (DSM) . . . . .	197
9.6	Raster Functions . . . . .	197
9.7	Local . . . . .	197
9.8	Focal . . . . .	198

9.9 Global . . . . .	201
9.10 Zonal . . . . .	201
9.11 Derivatives of Elevation Models . . . . .	202
9.12 Slope . . . . .	202
9.13 Aspect . . . . .	202
9.14 Heat Load Index . . . . .	205
9.15 Hillshade . . . . .	207
9.16 Sinks, Peaks, and Saddles Oh My! . . . . .	209
9.17 Landform Classification . . . . .	211
9.18 Profile and Planform Curvature . . . . .	211
9.19 Topographic Position Index . . . . .	214
9.20 Hydrology Work"flows" . . . . .	214
9.21 Flow Direction and Flow Accumulation . . . . .	215
9.22 Stream Delineation . . . . .	218
9.23 Stream Order . . . . .	219
9.24 Flow Length . . . . .	220
9.25 Watershed Delineation . . . . .	220
9.26 Topographic Wetness Index . . . . .	221
9.27 Case Study: Topographic Indices for Wetland Mapping . . . . .	222
9.28 DEM Derivatives and Classification . . . . .	226
9.29 3D Geovisualization . . . . .	226
9.30 Anaglyphs . . . . .	227
9.31 Viewsheds . . . . .	227
9.32 Extrusion . . . . .	228
9.33 Exaggeration . . . . .	228
9.34 Summary . . . . .	229
Reflection Questions . . . . .	229
Practice Questions . . . . .	229
Recommended Readings . . . . .	231
<b>10 Spatial Estimation</b> . . . . .	<b>233</b>
Learning Objectives . . . . .	233
Key Terms . . . . .	234
10.1 Introduction . . . . .	234
Recall This . . . . .	235
10.2 Geostatistics . . . . .	235
10.3 Spatial Autocorrelation . . . . .	235
10.4 Sampling . . . . .	235
10.5 Prediction . . . . .	235
10.6 Estimation . . . . .	235
10.7 Classical vs. Geostatistical Inferences . . . . .	235
10.8 Sampling . . . . .	236
Recall This . . . . .	237
10.9 Population . . . . .	237
10.10Sampling Design . . . . .	237
10.11Sampling Unit . . . . .	237

10.12Probability Sampling . . . . .	237
10.13Simple Random Sampling . . . . .	237
10.14Stratified Random Sampling . . . . .	238
10.15Systematic Sampling . . . . .	239
10.16Cluster Sampling . . . . .	241
10.17Non-probability Sampling . . . . .	241
10.18Representative Sampling . . . . .	243
10.19Unique Case Sampling . . . . .	243
10.20Sequential Sampling . . . . .	243
10.21Spatial Autocorrelation . . . . .	243
Recall This . . . . .	244
10.22Domain . . . . .	244
10.23Attributes . . . . .	244
10.24Moran's I . . . . .	244
10.25Case Study: Title of Case Study Here . . . . .	245
Calculating Moran's I . . . . .	246
Using Contiguity . . . . .	246
10.26Geary's C . . . . .	250
10.27Semivariogram Modeling . . . . .	252
10.28Case Study: Title of Case Study Here . . . . .	254
10.29Spherical . . . . .	256
10.30Exponential . . . . .	257
10.31Circular . . . . .	258
10.32Spatial Interpolation . . . . .	260
10.33Case Study: Title of Case Study Here . . . . .	260
10.34Methods Without Using Semi-variogram . . . . .	262
10.35Nearest Neighbor . . . . .	262
10.36Thiessen Polygon . . . . .	265
10.37Methods Using Semi-variogram . . . . .	268
10.38Kriging . . . . .	268
10.39Case Study: Title of Case Study here . . . . .	270
10.40Simple Kriging . . . . .	270
10.41Ordinary Kriging . . . . .	276
10.42Universal Kriging . . . . .	278
Which method is the best given our data? . . . . .	280
10.43Co-Kriging . . . . .	280
10.44Non-Linear Kriging . . . . .	281
10.45Indicator Kriging . . . . .	281
10.46Probability Kriging . . . . .	281
10.47Disjunctive Kriging . . . . .	281
10.48Spatial Regression Models . . . . .	282
10.49Case Study: Title of case study here . . . . .	282
10.50Spatial Lag Model . . . . .	282
10.51Steps in Fitting Spatial Lag Model: . . . . .	283
10.52Spatial Error Model . . . . .	285
10.53Steps in Fitting Spatial Error Model: . . . . .	286

<b>CONTENTS</b>	<b>9</b>
10.54Selection Between Lag and Error Model . . . . .	287
Remember This? . . . . .	288
Reflection Questions . . . . .	288
<b>11 Fundamentals of Remote Sensing</b>	<b>289</b>
Learning Objectives . . . . .	289
Key Terms . . . . .	289
11.1 What is Remote Sensing? . . . . .	290
11.2 Measuring Energy . . . . .	293
11.3 Introduction . . . . .	293
11.4 Electromagnetic Spectrum . . . . .	293
Call Out . . . . .	294
11.5 Scientific Notation . . . . .	295
11.6 Radiation Types . . . . .	295
11.7 Factors Affecting Radiation . . . . .	296
<b>11.8 Radiation Basics</b> . . . . .	296
11.9 Foundations of Measurement . . . . .	301
11.10Methods of Normalization . . . . .	301
11.11The Four Resolutions . . . . .	302
11.12Spatial Resolution . . . . .	302
11.13Temporal Resolution . . . . .	303
11.14Spectral Resolution . . . . .	303
11.15Radiometric Resolution . . . . .	304
11.16Key Applications . . . . .	307
11.17Case Study: Optical Remote Sensing to Evaluate Land Cover in Canada . . . . .	308
11.18Summary . . . . .	309
11.19Reflection Questions . . . . .	310
<b>12 Remote Sensing Systems</b>	<b>311</b>
12.1 Optical System Basics . . . . .	313
12.2 Perspectives . . . . .	325
12.3 Aerial Perspective . . . . .	326
12.4 Nadir and Zenith Perpsectives . . . . .	327
12.5 Oblique Perspective . . . . .	327
12.6 Hemispherical Perspective . . . . .	329
12.7 Platforms . . . . .	332
12.8 Orbital Physics . . . . .	339
Recall This . . . . .	342
12.9 Summary . . . . .	344
Reflection Questions . . . . .	344
Practice Questions . . . . .	344
<b>13 Image Processing</b>	<b>345</b>
Learning Objectives . . . . .	345
Key Terms . . . . .	345

13.1 Overview . . . . .	346
13.2 Geometric Correction . . . . .	346
13.3 Orthoimagery . . . . .	347
13.4 Relief Displacement . . . . .	348
Your turn! . . . . .	348
13.5 Georeferencing . . . . .	348
13.6 Georegistration (georectification) . . . . .	349
13.7 Resampling . . . . .	349
13.8 Nearest Neighbor . . . . .	350
13.9 Bilinear Interpolation . . . . .	350
13.10 Cubic Convolution . . . . .	350
13.11 Atmospheric Correction . . . . .	350
13.12 Atmospheric Windows . . . . .	351
13.13 Clouds and Shadows . . . . .	351
13.14 Smoke and Haze . . . . .	353
13.15 Radiometric Correction . . . . .	353
13.16 Signal-to-noise . . . . .	353
13.17 Readout Noise . . . . .	354
13.18 Thermal Noise . . . . .	354
Your turn! . . . . .	354
13.19 Case Study: Title of Case Study Here . . . . .	354
An overview of Landsat Processing . . . . .	354
13.20 Image Enhancement . . . . .	356
13.21 Stretching . . . . .	356
13.22 Smoothing . . . . .	356
Call out . . . . .	357
13.23 Summary . . . . .	357
Reflection Questions . . . . .	357
<b>14 Image Analysis</b>	<b>359</b>
Learning Objectives . . . . .	359
Key Terms . . . . .	359
14.1 Aerial Photography and Photogrammetry . . . . .	360
14.2 Image Classification . . . . .	360
14.3 Time Series Analysis . . . . .	360
14.4 Case Study: Sea Ice Change Analysis in the Beaufort Sea . . . . .	360
14.5 Pattern Analysis . . . . .	361
<b>15 LiDAR Acquisition and Analysis</b>	<b>367</b>
Learning Objectives . . . . .	367
Key Terms . . . . .	367
15.1 What is LiDAR? . . . . .	368
15.2 How Does LiDAR Work? . . . . .	368
15.3 LiDAR History and Use . . . . .	370
15.4 Components of a LiDAR System . . . . .	371
15.5 Lasers . . . . .	371

15.6 Position and Orientation . . . . .	374
15.7 Global Navigation Satellite Systems . . . . .	374
15.8 Inertial Measurement Unit (IMU) . . . . .	375
15.9 Clocks . . . . .	375
15.10 Platform . . . . .	375
15.11 Airplanes and Helicopters . . . . .	375
15.12 Drones . . . . .	376
15.13 Mobile Laser Scanning . . . . .	376
15.14 Satellite . . . . .	377
15.15 Types of LiDAR . . . . .	377
15.16 Discrete Return . . . . .	377
15.17 Full Waveform . . . . .	379
15.18 Emerging Technology . . . . .	380
Call Out . . . . .	380
15.19 LiDAR Derivatives and Analysis . . . . .	380
15.20 Bare Earth Elevation . . . . .	380
15.21 Digital Surface Model and Canopy Height Models . . . . .	381
15.22 Area Based Approach vs. Individual Tree Crown Approach . . . . .	382
15.23 Tree Segmentation . . . . .	383
15.24 Sources of Error . . . . .	384
15.25 Software and Analysis Tools . . . . .	384
15.26 Case Study: Creating LiDAR Metrics from a Raw Point Cloud . . . . .	386
Your Turn! . . . . .	387
15.27 Summary . . . . .	389
Reflection Questions . . . . .	389
Practice Questions . . . . .	389
Recommended Readings . . . . .	389
<b>16 Data Integration . . . . .</b>	<b>391</b>
Learning Objectives . . . . .	391
Key Terms . . . . .	391
16.1 Problems with Data Integration . . . . .	391
16.2 Framing The Problem . . . . .	392
16.3 About The Data . . . . .	392
16.4 Data Resolution . . . . .	394
16.5 Integrating Vector and Raster Data . . . . .	394
16.6 Rasterization . . . . .	394
16.7 Vectorization . . . . .	394
16.8 Zonal Statistics . . . . .	394
16.9 Smoothing . . . . .	395
16.10 Simplifying . . . . .	395
16.11 Spatial Data Errors . . . . .	395
16.12 Accuracy vs. Precision . . . . .	395
16.13 Vagueness and Ambiguity . . . . .	395
16.14 Quantifying Spatial Errors RMSE, Euclid's Distance . . . . .	396
16.15 Logical Errors . . . . .	396

16.16Ecological Fallacy, Atomistic Fallacy, MAUP etc. Its important to include these, whether here or elsewhere? . . . . .	396
16.17Other Errors? . . . . .	396
16.18Case Study: Title of Case Study here . . . . .	396
<b>17 Making Beautiful Maps</b>	<b>401</b>
Learning Objectives . . . . .	403
Key Terms . . . . .	403
17.1 Types of Maps . . . . .	403
17.2 Thematic Maps . . . . .	403
17.3 Choropleth Maps . . . . .	407
17.4 Dot Density Maps . . . . .	408
17.5 Isoline Maps . . . . .	408
17.6 Diagrammatic Maps . . . . .	408
17.7 Cartograms . . . . .	408
17.8 Additional Resources on Types of Maps . . . . .	408
17.9 Map Composition . . . . .	410
17.10Figure . . . . .	410
17.11Ground . . . . .	410
17.12Frame . . . . .	411
17.13Elements of Maps . . . . .	411
17.14Text . . . . .	411
17.15Legend . . . . .	411
17.16Scale and North Arrow . . . . .	413
17.17Measured Grid . . . . .	413
17.18Citation . . . . .	413
17.19Symbolization . . . . .	413
17.20Separable . . . . .	414
17.21Integral . . . . .	414
17.22Graduated . . . . .	414
17.23Configurable . . . . .	415
17.24Proportional . . . . .	415
17.25Line Weight . . . . .	415
17.26Additional Resources . . . . .	415
17.27Colour . . . . .	415
17.28Hue . . . . .	417
17.29Chroma . . . . .	419
17.30Lightness . . . . .	419
17.31Bivariate Colour Schemes . . . . .	419
17.32Colour Pickers . . . . .	419
17.33Additional Resources . . . . .	420
17.34Classification Schemes . . . . .	420
17.35Qualitative . . . . .	420
17.36Sequential . . . . .	420
17.37Intervals . . . . .	420
17.38Quantiles . . . . .	420

17.39Natural Breaks (Jenna) . . . . .	420
17.40Standard Deviation . . . . .	420
17.41Additional Resources . . . . .	420
17.42Generalization . . . . .	421
17.43Select . . . . .	421
17.44Amalgamate . . . . .	421
17.45Exaggerate . . . . .	421
17.46Displace . . . . .	421
17.47Refine . . . . .	421
17.48Simplify . . . . .	421
17.49Aggregate . . . . .	421
17.50Typify . . . . .	421
17.51Smooth . . . . .	421
17.52Enhance . . . . .	421
17.53Collapse . . . . .	421
17.54Merge . . . . .	421
17.55Additional Resources . . . . .	421
17.56Map Design . . . . .	422
17.57Subject . . . . .	422
17.58Projection and Orientation . . . . .	422
17.59Hierarchy . . . . .	422
17.60Balance . . . . .	422
17.61Summary . . . . .	422
Reflection Questions . . . . .	422
Practice Questions . . . . .	423
Recommended Readings . . . . .	423



# About this textbook

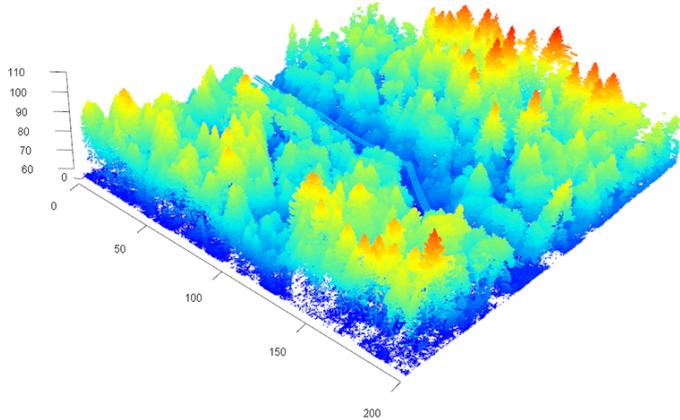


Figure 1: Cover design by Lily Crandall-Oral, CC-BY-SA-4.0.

Geomatics is a large and growing field that encompasses the art, science and technology of measuring attributes about Earth's systems. From the earliest observations of Earth's shape in ancient Greece to the space and information age, humans have wondered about their place on this planet. Nearly every aspect of our modern geographic lives can be attributed to advances in the field of geomatics. Today, location is real-time and precise, made easy by millennia of observations and incremental advances in technology that have culminated in a wondrous field of study.

The purpose of this textbook is to give students and practitioners a solid survey (pun intended) of what modern geomatics is capable of when confronting environmental management problems. We take a Canadian perspective to this approach, by telling the historical contributions of Canadians to the field and sharing real-world case studies of environmental management problems in Canada. Inside, you will find interactive web-based content and tools, case studies with real geographic data, and tons of reflection and practice questions. Best of

all, unless otherwise stated, the content is free to use, adapt, and remix for any purpose with attribution to us and under the same Creative Commons 4.0 license.

## Who Should use this textbook

This textbook is designed to be used by students, instructors, and practitioners of geomatics. All the basics are covered for introducing Geographic Information Systems (GIS) and remote sensing at any level or as a refresher. We throw in a little extra for understanding common GIS workflows and some more advanced topics that are suitable for advanced students. Unlike many other geomatics textbooks, we do not adopt any specific GIS software and therefore any student or practitioner should be able pick up this textbook and apply the methods and knowledge to the software of your choice. That said, the textbook is created with many geospatial packages in R and because it is open, anyone is free to poke their head under the hood in the GitHub repository for specific workflows and geospatial data handling with R.

## How to adopt this textbook in your geomatics classroom

Depending on what you are teaching and how you organize your course, you can easily assign readings that are thematic or sequential. A great way to get started with adopting this textbook is to read the clear and concise **Learning Objectives** that are displayed at the very beginning of each chapter and then compare with your own curriculum.

### LEARNING OBJECTIVES #

1. Understand what a Geographic Information System (GIS) is and what it can do
2. Understand how the five components of GIS are integrated to solve spatial questions
3. Recognize the role of models and maps in our understanding and representation of location
4. Recognize the significant Canadian contributions in the advancement of geomatics as a field

Chapter 1 provides an overview of the field with some key definitions and a detailed history of much of the technology that is presented in later chapters. The remainder of the textbook can be divided into four parts:

- **Part 1** Chapters 2-5 cover **Geospatial Data Fundamentals** such as types of data, collecting and editing data, Global Navigation Satellite Systems, and relational database management systems.
- **Part 2** Chapters 6-10 focus on **Spatial Analysis** and cover a wide variety of common workflows, some of which might be found in more advanced geomatics coursework like network analysis and spatial estimation.

- **Part 3** Chapters 10-15 focus on everything **Remote Sensing** including theory, systems, image processing and analysis, and LiDAR (light detection and ranging).
- **Part 4** Chapters 16 and 17 illustrate how to create **Information from Data** with a focus on integrating data and creating maps.

Chapters are written to stand alone as authoritative contributions on specific topics, but an accumulative approach is taken to content organization. In this way, the textbook can be adopted in sequence as-is for an entire semester or you can adopt the chapters that suit the needs for your course and compliment them with other materials. Each chapter is also subdivided into subsections that can be easily linked to for an assigned reading.

## How to get involved

Because this is an open project, we highly encourage contributions from the community through our **Open Geomatics Community of Practice** on GitHub. The textbook is hosted on our GitHub repository and from there you can open an issue or start a discussion. Feel free to open an issue for any typos, factual discrepancies, bugs, or topics you want to see in a future edition. We are always looking for great Canadian case studies to share! Beyond the built versions of the textbook (e.g., [[www.opengeomatics.ca](http://www.opengeomatics.ca)(<http://www.opengeomatics.ca>)], pdf, epub, etc.), you can fork our GitHub repository to explore the source code.

## Acknowledgements

So many folks helped with the development of this textbook and many of them are named as authors in various chapters. For those not explicitly named with contributions, we greatly thank the efforts by Nick Murphy for collating our bibliographic sources, Natasha Sharma for editing and fixing things, students enrolled in the Master of Geomatics for Environmental Management program at the University of British Columbia for contributing case studies and early reviews. This textbook would not have been possible without the support of an Open Educational Resources (OER) Implementation Grant from the University of British Columbia.



# **Chapter 1**

## **What is Geomatics?**

Written by Paul Pickell

We encounter and use geographic information on a regular basis in our everyday lives. Whether it is finding directions to a retailer that has an item in stock you want to buy or recording the path of your last morning jog, you have probably used a Geographic Information System (GIS) and not even realized it. In this chapter, we explore what geomatics is all about.

### **Learning Objectives**

1. Understand what a Geographic Information System (GIS) is and what it can do
2. Understand how the five components of GIS are integrated to solve spatial questions
3. Recognize the role of models and maps in our understanding and representation of location
4. Recognize the significant Canadian contributions in the advancement of geomatics as a field

### **Key Terms**

Geomatics, Geographic Information Systems (GIS), Remote Sensing, Capture, Store, Analyze, Model, Data Model, Conceptual Model, Map, Query, Analyze, Display, Output

## 1.1 The Science and Technology of Geomatics

**Geomatics** is the science and technology of collecting geographic data and converting it to geographic information for use in a wide variety of industries. As a technical field, it encompasses many different work processes including surveying, remote sensing, global navigation satellite systems, geospatial analysis, and information technology and systems management. In turn, these processes support a wide variety of spatial decision-making such as urban planning, ecological conservation, forest management, real-time planetary systems monitoring, and rapid response to natural disasters. Many more emerging technologies such as self-driving vehicles, ride-sharing apps, and augmented reality video games depend directly on the science and technology of geomatics.

## 1.2 Information Systems

An **information system** is used to store, code, and recall information. In the Information Age, we are surrounded and depend heavily on information systems such as financial systems that record the transactions in your bank account or navigation systems that tell you the fastest route to a destination or autonomous vehicles such as the SkyTrain rapid transit network in Vancouver, Canada that moves more than a half million people every day across the Metro Vancouver region. These are all examples of systems that require high synchronization and integration of many varied sources of information in order to move people and assets around. It should come as no surprise then that information systems and information technology contribute significantly to nearly every sector of developed and developing economies.

What makes a **Geographic Information System** different from other information systems is the type of information that is handled: geographic location. With a GIS you can know the quantity and quality of something and the **location** of that event, activity, or feature. For example, you might have recorded your heart rate and timed your morning jog using your phone or other fitness tracking device. Since you also have the location or coordinates of your jog, you can calculate or derive additional information such as speed (distance per time), the total distance you jogged, and also your elevation above sea level. You could even relate your heart rate to different locations along your jog to better understand your performance on different terrain. This is an example of a GIS at work. You are storing geographic information into a system that allows you to code the information with different qualities (e.g., the type of surface you ran on) and quantities (e.g., your heart rate) and then recall that information in a way that allows you to explore trends and ask and answer **spatial** questions.

## 1.3 The Five Components of GIS

A GIS is a collection of five inter-connected, inter-acting, and inter-dependent components (Figure 1.1).

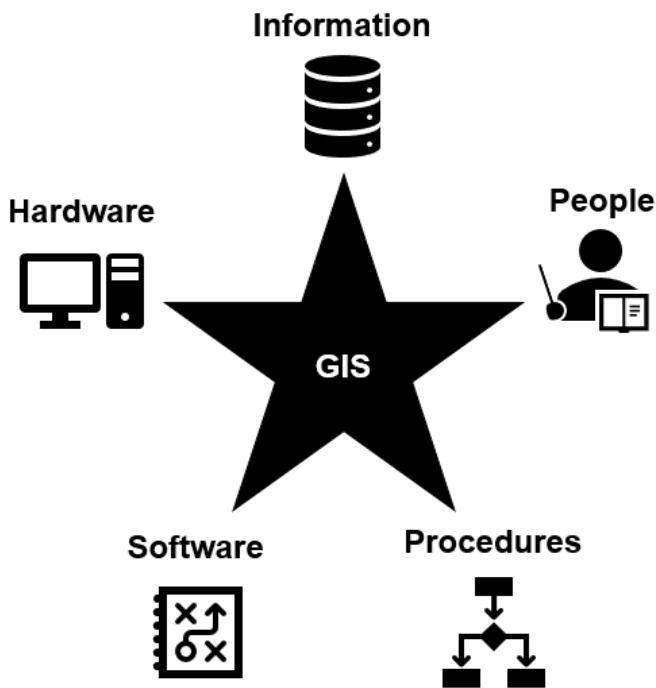


Figure 1.1: The five components of a geographic information system. Pickell, CC-BY-SA-4.0.

**Data** – Data are a collection of qualitative and quantitative variables about real world entities. At the very least, the type of data stored in a GIS is usually spatial in nature, meaning we have some set of coordinates or representation of location. But a GIS can also store aspatial data, data that does not necessarily have coordinates, but may still be relevant to other geographic data. For example, you might also record the dates that you jogged or the weather during your jogs. These are examples of aspatial information that we might want to relate to your jogs.

Spatial data can take many forms in a GIS, which we will explore more in Chapter 2. For the sake of our jogging example, you can imagine your jogging path could be represented as a line or a set of ordered points. In turn, your jogging path can be described by different qualities (day of the week, weather) and quantities (distance, speed, heart rate), which are called **attributes** (Figure 1.2). An important role of a GIS is to govern how data can interact and behave

### Attribute table

ID	Location	Heart rate	Surface type
1	$(x_1, y_1)$	200	Pavement
2	$(x_2, y_2)$	210	Pavement
3	$(x_3, y_3)$	205	Trail

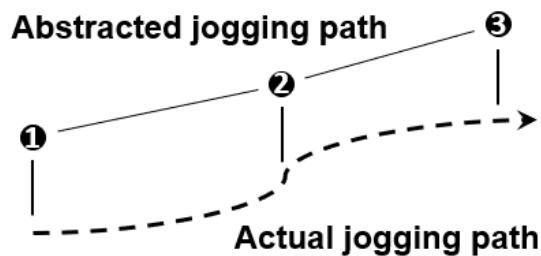


Figure 1.2: Abstracted spatial and aspatial information from a jogging path represented by three ordered points. Pickell, CC-BY-SA-4.0.

with other types of data. For example, we can calculate distance from a line, but not area, which is a special property of a polygon. We will explore data behaviors in more detail in Chapter 2.

**Software** – There are a lot of different GIS software packages available and the choice of which software to use is usually driven by preference, availability, and/or cost. For example, some GIS software are freely available while others require a paid license. Nearly all GIS software packages can perform the same operations and analyses, but there are some differences like integrated cloud computing or availability of certain plugins. The most widely-used paid software is ESRI's ArcGIS, while the most widely-used Free and Open Source Software (FOSS) is QGIS (Q for Quantum). There are many other historically-significant FOSS software packages like GRASS, SAGA, and MOSS that are discussed in more detail later in this chapter.

**Hardware** – GIS software operate on hardware comprised of computer components and human interface devices (HID). High resolution and high refresh rate monitors are often required when dealing with high resolution imaging and other digitizing, which require high-end graphics cards. Solid State Disks (SSD) with high capacity are important for fast read-write speeds to access and process data while Random Access Memory (RAM) and Central Processing Unit (CPU) clock speed can also be limiting factors for processing. Some specialized HID are designed specifically for working with GIS such as active shutter and

polarized glasses for stereo vision, and joysticks for navigating in 3D space.

**Methods** – All organizations must design business processes in order to effectively operationalize GIS for solving real-world problems. Methods will vary from organization to organization based on industry, client needs, and the types of data needed. For example, the business process of a forestry company will involve digitizing forest stands and other natural resources, enumerating attributes for stand inventories, road and forest harvest planning, forest regeneration monitoring, and creating operational maps. It is important for an organization to define exactly the activities, standards, and quality assurance (i.e., methods) that will be needed to achieve operational goals. Well-defined and well-documented methods also ensure that organizational knowledge is retained during staff turnover.

**People** – We are responsible for identifying the information needs, defining the procedures, coding the software, and building the hardware for a Geographic Information System. We are also responsible for interpreting, analyzing, and reporting our results and communicating our findings through maps and reports. Geographic Information Systems do not occur naturally and because they are a product of our imagination and skill, they inherit our best and worst traits. For example, we often make mistakes when we define our procedures or identify our information needs, which lead to incorrect conclusions. Moreover, our biases intentionally and unintentionally creep into our map-making, distorting shape, size, and importance of features that we want to represent.

## 1.4 What a GIS can do

Like other information systems, a GIS should be able to do the following things with geospatial data:

- Capture geospatial data
- Store geospatial data
- Query geospatial data
- Analyze geospatial data
- Display geospatial data
- Output geospatial data

Geospatial data may be **captured** from remote sensing (Chapter 11) or digitized/scanned from other sources (Chapter 4). Data can be **stored** in many different formats with different behaviours and rules and are kept in databases (Chapter 3). Data can be queried (or selected) from a database using a standard query language (Chapter 5). Once data are queried, they can then be **analyzed** using operations like proximity, overlay, intersection and zonal statistics (Chapter 6). A primary justification for investing resources into a GIS is to display and **output** data in the form of graphs, tables, charts, and maps (Chapter 17).

A GIS ultimately answers spatial questions like:

- Where should I buy a home?
- Where is population density highest in Vancouver, Canada?
- Where is the most economical location for a new mill?
- Where should land be protected or conserved?

## 1.5 Modelling the world with GIS

Together, the five components of a GIS allow us to model our spatial environment. A **model** is an abstraction or simplification of reality. Models are necessary for us to understand how complex things work and communicate that understanding to others. For example, the hydrologic cycle describes the states and processes of water in the atmosphere, over land, in the ocean as well as forms of precipitation and evaporation. The concept of the hydrologic cycle is a simplification, of course, because movement and the phase of water is caused by all kinds of interacting forces such as gravity, temperature, pressure, and climate. For this reason, the hydrologic cycle is an example of a **conceptual model**, which is used to hypothesize and theorize about how our world works. This brings us to the second important quality of models: they are reproducible. A model must be reproducible so that we can validate it or check that the result or output is in line with what we expect or observe. For example, the principles of the hydrologic cycle are universal, that is they apply anywhere on Earth and even on other planets because they are based on physical properties that are shared across our universe (e.g., temperature, pressure, gravity).

We can also create models that allow us to play out different scenarios and assess the possible range of outcomes in geographic space. For example, you might be interested in purchasing a home, but you have some spatial and aspatial criteria such as price range, preference of neighborhood, distance to your work, or the quality of the nearby schools. This is the type of spatial optimization problem that we often seek to solve with a GIS. If you go onto a property listing website, you will notice that the results will change based on the criteria that you enter. And this brings us to the final type of model that is important to a GIS: the map.

**Maps** are spatial models. They abstract geographic information like location and attributes of things. It is important to underscore the difference between spatial models such as maps and conceptual models such as the hydrologic cycle. With a map, you could abstract or represent the location, distance and area of a lake, but you would need the conceptual model of the hydrologic cycle to describe the interaction of the lake with the atmosphere (i.e., evaporation). There are many aspects of the hydrologic cycle that can be specifically modelled in a GIS, which is something we will explore further in Chapter 9.

## 1.6 A Canadian History of GIS

Modern GIS cannot be understood without the context of three major technological leaps in the nineteenth and twentieth centuries: the photograph, the airplane, and the computer. None of these technologies were discovered by Canadians, yet all were instrumental for the Canadian cartographic advances that followed.

### 1.6.1 The Photograph

The first successful attempt to record light on a durable image can be traced to France. Nicéphore Niépce is widely credited with the invention of photography by producing the oldest surviving photograph from the window of his home in 1826 or 1827 simply titled, *View from the Window at Le Gras* (Figure 1.3). His process involved using a camera obscura (a discovery handed down from antiquity) to project a real-world scene onto a photosensitive metal plate for hours and possibly days on end. Though the results were crude, it was a major advancement over all earlier attempts that failed to produce a durable photograph. Niépce died only a few years later in 1833 before he was able to publish his invention. Before his death, he shared his secrets with Louis Daguerre who would perfect his process and announce the daguerreotype to the world in 1839, a photographic process that was much faster and commercially viable. The announcement of the daguerreotype from France spread rapidly and resulted in the first known photograph captured of Canada the very next year of Niagara Falls, Ontario (Figure 1.4).

It was not long before more Canadian landscapes became the subjects of photography. In the next decade, French colonel Aimé Laussedat developed a new instrument that combined a contemporaneous camera with a theodolite (a surveying instrument) that he called a phototheodolite. The phototheodolite allowed photographs to be precisely taken such that multiple photographs could provide various perspectives of known locations in order to produce topographic maps. By 1867, Laussedat exhibited the first map of Paris produced with a phototheodolite and in doing so he ushered in the field of **photogrammetry**, the science of deriving 3D measurements from 2D photographs. In 1887, the Geological Survey of Canada set out to photograph the Rocky Mountains using these photogrammetric techniques and the Canadian Rockies became one of the most photographed landscapes of the day with over 25,000 historical images captured between 1887 and 1958. Today, these historical images have been digitised from Library and Archives Canada through the Mountain Legacy Project and many locations have been re-imaged more than a century later by researchers at the University of Alberta and University of Victoria, illuminating dramatic changes to glaciers and forest cover on these Canadian landscapes (Figure #).



Figure 1.3: View from the Window at Le Gras showing some buildings and a tree in the distance. This image has been flipped from the original metal plate along both the horizontal and vertical axes. The camera obscura would have originally recorded the upper right corner in the lower left corner. Public Domain.



Figure 1.4: The earliest known photograph of present-day Canada was this daguerreotype taken in 1840 by Hugh Lee Pattinson of Niagara Falls, Ontario (Horseshoe Falls). Due to the daguerreotype capture process, the image is flipped on the vertical axis so that image right is the United States on the East side of the falls and image left is Canada on the West side of the falls (image is looking South). Courtesy of Newcastle University Library Special Collections, CC-BY-SA-4.0

### 1.6.2 The Airplane

The next technological leap came at the turn of the century and forever transformed photography and photogrammetry. From 1900 to 1902, Orville and Wilbur Wright had been designing and experimenting with light-than-air gliders and kites near Kitty Hawk, North Carolina. By 1903, the brothers had designed the first heavier-than-air airplane capable of powered flight by a gasoline engine called the *Wright Flyer*. Incidentally, because the technology of photography had developed six decades earlier, the moment of this first flight was captured in one of the most iconic photographs ever taken by John T. Daniels, member of the U.S. Lifesaving Service Station at Kill Devil Hills who was standing by (Figure 1.5).



Figure 1.5: Restored image taken by John T. Daniels on December 17, 1903 near Kitty Hawk, North Carolina only seconds into the 12 second first powered flight piloted by Orville Wright. Wilbur Wright is seen to the right. Public Domain.

Aviation developed quickly following the first successful powered flight by the Wright Brothers and so did aerial photography. In 1912, Frederick Charles Victor Laws of the British Royal Flying Corps discovered that aerial photographs taken with at least 60% overlap could be used to produce a 3D stereographic effect when viewed through a **stereoscope** (Figure 1.6). Following Laws' realization, flight paths were more intentionally planned to achieve this effect and in the coming years many more personnel were employed to help with **air photo interpretation**, the process of viewing air photos and discerning types of features based on tone, shade, shape, pattern, texture, and spatial association. At the outset of World War I, aircraft were being equipped with cameras for reconnaissance at the front lines.



Figure 1.6: A woman is seen annotating an overlapping air photo pair with a stereoscope in 1945. Canada. Dept. of National Defence / Library and Archives Canada / PA-065599, Public Domain.

The Canadian government quickly realized the potential benefits of air travel and aerial surveying for such a large country. In 1919, Canada became the first nation to regulate the domain of aviation with the establishment of the Air Board. Six air stations were established by the Air Board around the country primarily for civil patrols and surveying (Figure 1.7). By 1922, Canada established the National Air Photo Library to store the new and increasing cache of air photos being collected by federal agencies. During the 1930's, many private companies began prospecting from the sky and found profit in undertaking aerial mineral surveys in the Northwest Territories and the Yukon.

Airplanes played a more significant role in World War II both as weapons and for reconnaissance. A new process for collecting aerial imagery was developed called **trimetrogon**, which involves capturing three photos at the same time: one directly down and two oblique images to the right and left of the airplane (Figure 1.8). Even in the 1940's, much of Canada remained unmapped and unimaged, so the Royal Canadian Air Force contributed multiple photo squadrons with the task of mapping the vast country. One of those squadrons was No. 413 Photographic Squadron (Figure 1.9), which was based out of Canadian Forces Base Rockcliffe near Ottawa (formerly Rockliffe Air Station, the first air station established in Canada) and responsible for capturing the first aerial images of northern Canada. In the summer of 1950, No. 413 Photographic Squadron completed imaging the last gaps of land near Ellesmere Island. For the first time ever and more than 30 years after aerial surveys began, Canada finally had a self portrait from the sky. Today, a century after aerial surveying efforts began, the National Air Photo Library archives over six million air photos.



Figure 1.7: One of the first air stations established after the inception of the Air Board was Jericho Air Station seen here in 1923 near Vancouver, Canada. Canada. Dept. of National Defence / Library and Archives Canada / PA-051943, Public Domain.



Figure 1.8: Trimetrogon camera installation in North American "Mitchell" II aircraft of No. 14 Photo Squadron, Royal Canadian Air Force seen in 1945. One of the oblique angle cameras can be seen on the side of the aircraft toward the nose. Library and Archives Canada / PA-065503, Public Domain.



Figure 1.9: Inside the fuselage of an aircraft of RCAF No. 413 Squadron showing an alternative arrangement of trimetrogon cameras with the oblique cameras pointing inward instead of outward. Library and Archives Canada / PA-065920, Public Domain.

After the Second World War, many RCAF pilots went on to start their own aerial surveying companies and benefited from the immense mapping efforts that were underway across the nation. One particular company, Spartan Air Services, emerged in 1946. Founded by two retired RCAF pilots, Spartan Air Services primarily flew aerial surveys for

### 1.6.3 The Computer

### 1.6.4 The Land Use Problem in Canada

1935 - Prairie and Farm Rehabilitation Act to address drought, farmer insolvency, and unemployment. Dust bowl.

The post-war economy of Canada saw significant growth in the 1940s and 1950s and huge demographic shifts were underway. Agrarian societies were transformed by the new economy and rapid urbanization. Prior to the Second World War, about half of all Canadians lived in rural areas (?). By the 1950s, nearly 62% of the total population lived in urban centres and 37% lived in just the fifteen largest cities (Figure 1.10, of Statistics [1952]). Faced with the prospect of failed crops and uncertain economic outcomes, many chose to head for better opportunities promised in the metropolitan areas. In the decade from 1941 to 1951, 105 Canadian families left their farms every week, despite an overall increase in farmland across the country (of Statistics [1952], of Statistics [1953]).

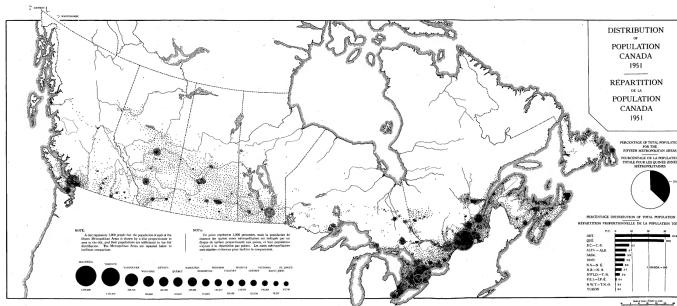


Figure 1.10: Ninth Census of Canada 1951, Population Distribution. Dominion of Bureau Statistics, Public Domain.

As cities sprawled, land use and land planning conflicts emerged. City waistlines grew as adjacent farmland was consumed to welcome new urban dwellers. In 1957, the Senate established the Special Committee on Land Use in Canada whose purpose was “to consider and report on land use in Canada and what should be done to ensure that our land resources are most effectively utilized for the benefit of the Canadian economy and the Canadian people and, in particular, to increase both agricultural production and the incomes of those engaged in it” (The Senate of Canada [1957]). Over the next four years, the special committee heard testimony from 109 witnesses and recorded 1,606 pages of evidence on the issue of land use in Canada (Special Committee on Land Use in Canada [1963]).

Following years of hearings by the Senate Special Committee on Land Use in Canada, the Agricultural and Rural Development Act was passed by Parliament in June 1961 (McCrorie [1969]). The new legislation empowered the Minister of Agriculture to work with the Provinces and Territories to address rural poverty and promote the development and conservation of Canada’s prized farmland. By October, the federal government had organized a national conference in Montreal, “The Resources for Tomorrow”. Over 700 delegates from government, non-governmental organizations, university, and industry contributed to a unique national conversation on the future of Canada’s renewable natural resources with workshops on agriculture, forestry, fisheries, recreation, wildlife, water and urban growth (Gray [1962]). The conference was attended by then Governor General Georges Philias Vanier, Prime Minister Diefenbaker, and future Prime Minister Pierre Elliot Trudeau among many other bureaucrats from all levels of government. The proceedings were highly spatial in nature with a total of 66 maps found among the 87 published background papers, all hand-drawn of course.

[CONCLUSIONS AND RECOMMENDATIONS FROM CONFERENCE HERE] It was clear that a national approach to mapping renewable resources was needed in order to protect and conserve them, so in 1963 the federal government established an ambitious project called the Canada Land Inventory

whose chief purpose was to provide a “comprehensive survey of land capability and use designed to provide a basis for resource and land use planning” (of Regional Economic Expansion [1965]). Initially, the Canada Land Inventory prioritized soil maps, a nod to the strategic importance of vanishing farmland

But how could the Canada Land Inventory ever be achieved for such a vast country using traditional paper maps? To be sure, Canada was not lacking the mapped data. In fact there were tens of thousands of maps produced, cataloged, and housed by the federal government on everything from soils to agriculture to recreation to forests. Still, the scale of agricultural land loss was not known at the time because paper maps of soil fertility could not easily be combined with other information like census data. Decision-makers relied on human interpretation and analysis of paper maps, a time-consuming task that could only accommodate a handful of mapped attributes at a time over a small area. Paper maps had to be reproduced at the same scale, aligned, and analysed polygon-by-polygon. The process to create maps at the time was entirely manual, requiring approximately 10 hours by a skilled technician to produce a single map sheet and about as much time by another to check and make any corrections (Tomlinson [1974]). At that rate, it was estimated to take 666 “man-years” and 133 staff to complete the project (?). A new approach was needed to cope with the overwhelming amount of information being produced.



Figure 1.11: Roger Tomlinson talks about the use of computers for the Canada Land Inventory in 1967. Frame from "Data for Decisions", a National Film Board of Canada documentary, directed by Michael Millar.

It was 29-year old Roger Tomlinson who recognized the need for computers to handle the massive influx of mapped data (Figure 1.11). Tomlinson was working in Ottawa as an air photo interpreter for the aerial survey company Spartan Air Services, a company he joined in 1959. The phenomenon that occurred next can only be explained by American geographer Waldo Tobler's **First Law of Geography**, “everything is related to everything else, but near things are

more related than distant things” (?). In spring 1962, aboard a flight bound for the capital city where Spartan Air Services was headquartered, Tomlinson serendipitously met the lead bureaucrat of the Canada Land Inventory, Lee Pratt. Pratt and Tomlinson exchanged ideas about creating a computer system for handling mapped data. Tomlinson outlined a possible approach to help automate the mapping approach using computers and with some napkin

Cartographic puberty, the awkward but necessary transition from paper to magnetic tape.

to lead the development of the world’s first computerized GIS. That was 1962. Tomlinson first described it as a “geo-information system” (Tomlinson [1967]), but it was later renamed the Canadian Geographic Information System before it was shortened to the **Canada Geographic Information System (CGIS)**. In the years that followed, the Canada Land Inventory produced over 12,000 map sheets showing the capability of Canada’s lands for agriculture, forestry, recreation, and wildlife (Fisher and MacDonald [1979]). At the time, a major function of the CGIS was to scan and digitize maps, all of which were on paper sheets.

Other advances in computer hardware were achieved as a result of the pioneering work by Tomlinson, including the development and integration of map scanners and digitizer tables for inputting and updating geographic data into the computer. Tomlinson is widely acknowledged as the creator of modern computerized GIS.

While the Canada Land Inventory was underway with the aid of the CGIS, provinces moved to create the agricultural land reserves to ensure that the most fertile areas remained available as farmland. In 1973, British Columbia established the Agriculture Land Reserves, setting aside 5% of its land base for farming priority. Newfoundland and Labrador established an Agriculture Development Area around the city of St. John’s in the same year. By 1977, it was learned from CGIS analysis that only 0.5% the second largest country in the world supports prime farmland and half of that occurs in the vicinity of the most populous metropolitan area: Toronto (Figure 1.13) (Manning and McCuaig [1977]). Ontario published Food Land Guidelines the next year that effectively laid out a land use planning framework for municipalities and the province to protect and maximize the efficiency of farming on the best available land (of Agriculture and Food [1978]).

On the opposite shore of Lake Erie from where some of the best Canadian soils for farming were deposited millennia ago by the glacial processes that formed the Great Lakes, the Cuyahoga River quietly empties a small drainage basin of approximately 2,100 km<sup>2</sup> and bisects the city of Cleveland, Ohio in the United States. This relatively small, crooked river is one of many that fills Lake Erie and flows slowly by former sleepy industrial towns along the lake shore. The first European settler built his cabin on the shore of the Cuyahoga River in 1797 at the site of present-day Cleveland and over the next 155 years of industrialization,

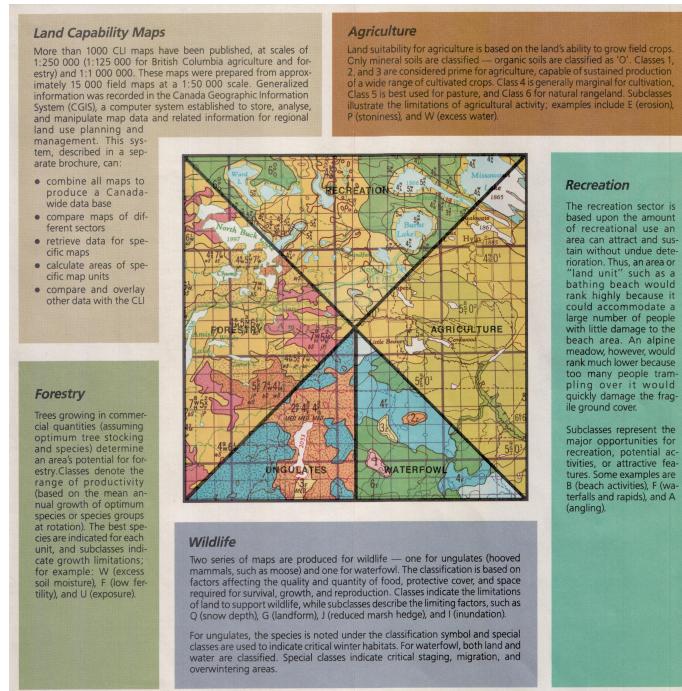


Figure 1.12: Excerpt from a brochure describing initial mapping efforts of the Canada Land Inventory in British Columbia. Published by the Minister of Supply and Services Canada, n.d.

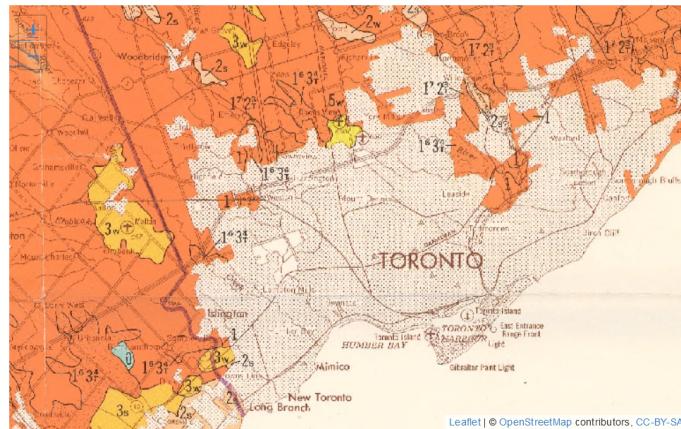


Figure 1.13: Soil capability for agriculture from the Canada Land Inventory for the area surrounding Toronto, Canada. The majority of the best soils for farming in Canada (Class 1 - No significant limitations in use for crops) are found in southeastern Ontario. Queen's Printer, Public Domain. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:1-cli-agriculture-capability-toronto>.

the river was reported to have caught fire nine times (La Bella [2009]). Ignited once every 17 years or so from volatile oil slicks and other pollutants suspended in the water column (Figure 1.14).

In 1968, 16 years after the last burning, just 60 km upstream, a symposium at Kent State University described the state of a 4 km reach of the river (University [1968]):

[...] Large quantities of black heavy oil floating in slicks, sometimes several inches thick, are observed frequently. Debris and trash are commonly caught up in these slicks forming an unsightly floating mess. Anaerobic action is common as the dissolved oxygen is seldom above a fraction of a part per million. The discharge of cooling water increases the temperature by 10 to 15 °F [5.6 to 8.3 °C]. The velocity is negligible, and sludge accumulates on the bottom. Animal life does not exist.

The following year in the summer of 1969, the Cuyahoga River burned again. This time, the fire ignited a national conversation in the United States about pollution and was a catalyst for the passage of landmark legislation for protecting the environment. A month after the fire was extinguished, the United States Senate unanimously passed the National Environmental Policy Act (NEPA), legislation that required the federal government to consider the environmental impact of its land use. The following year, the Environmental Protection Agency (EPA) was established by an executive order of President Nixon and



Figure 1.14: The Cuyahoga River burns in 1952 near Cleveland, Ohio in the United States as an oil slick is ignited. Special Collections, Cleveland State University Library.

the Clean Air Act was passed. By 1972, the United States Congress passed the Clean Water Act. These events in the late 1960s and early 1970s provided new environmental and regulatory imperatives for the continued development of GIS.

By 1976,

Landsat Island, a truly Canadian story of discovery:

I have spoken to Frank Hall Sr. and he told me a fascinating story about the moment of discovery. He was strapped into a harness and lowered from a helicopter down to the island. This was quite a frozen island and it was completely covered with ice. As he was lowered out of the helicopter a polar bear took a swat at him. The bear was on the highest point on the island and it was hard for him to see because it was white. Hall yanked at the cable and got himself hauled up. He said he very nearly became the first person to end his life on Landsat Island.

## 1.7 Summary

### Reflection Questions

1. What is the difference between a conceptual model and a spatial model?
2. What was the initial purpose for developing the first GIS?

### Practice Questions

1. Name the five components of GIS.
2. Give an example of a spatial question that could be answered by a GIS.

3. What is a model? Give an example of a model.
4. What is a map?
5. What was the name of the first computerized GIS? Who was involved in its development it?



# Chapter 2

# Mapping Data

Written by Paul Pickell

You probably already accept that the Earth is “round” and not “flat”. You have probably held and touched a globe at some point in your life. But have you ever wondered how we describe location and measure something as large as the Earth? In this chapter, we will explore fundamental concepts for how we measure the Earth and orient ourselves with coordinate systems.

## Learning Objectives

1. Understand the models of Earth’s figure and shape
2. Describe different vertical datums and how they are used to reference height
3. Understand the difference between cartesian, celestial, geographic, and projected coordinate systems
4. Recognize the differences among major types of map projections
5. Explore how projected coordinate systems distort and represent the world around us

## Key Terms

Antipode, Great Circle, Small Circle, Geodesy, Vertical Datum, Horizontal Datum, Deflection of the Vertical, Ellipsoid, Spheroid, Geoid, Elevation, Orthometric Height, Geoid Height, Geodetic Height, Coordinate System, Celestial Coordinate System, Cartesian Coordinate System, Geographic Coordinate System, Projected Coordinate System, Map Projection, Tissot’s Indicatrix

## 2.1 Introduction to Geodesy

**Geodesy** is the fascinating science of measuring the shape, orientation, and gravity of Earth. Naturally, some of the questions that come to mind when thinking about such a grand topic are *I thought the shape of Earth is a sphere?* and *How do we orient ourselves on Earth?* and *What does gravity have to do with mapping location?*

All of these questions stem from need to represent **location**. For our purposes, location is the position of something relative to something else. In order to actually describe a location on Earth, we first need to know the size and shape of Earth. Some of the first estimations of Earth's size and shape were made by Eratosthenes, a Greek mathematician from the second and third centuries B.C. Eratosthenes was responsible for many concepts we use in our everyday lives:

- Conceiving the first spherical model of Earth
- The first accurate measure of Earth's circumference
- Calculating the tilt of Earth's axis
- Calculating the distance of Earth to the Sun
- Invention of the leap day

Eratosthenes accurately calculated the circumference of Earth by noticing how the Sun shone directly down the bottom of a well in Syene (modern Egypt) at noon on the summer solstice. He later made a second observation at Alexandria at noon on the summer solstice with a pole and noticed a shadow. He measured the angle of the shadow and inferred the circumference of Earth, which was already known to be spherical (Figure 2.1).

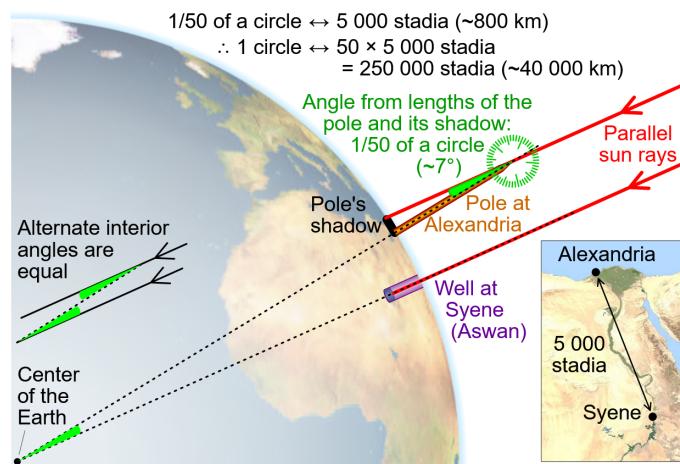


Figure 2.1: Diagram showing how Eratosthenes estimated the circumference of Earth by observing the angle of a shadow that was cast about 800 km north of Syene in present-day Egypt. [Monniaux et al., 2005], CC-BY-SA-4.0.

Pretty simple, right? Turns out, Eratosthenes was off by only 75 km or less than 0.2% in his calculation! The actual North-South circumference of Earth is about 40,075 km. His calculation worked because the Sun's rays are nearly parallel when they strike Earth. So if you observe the Sun at the same time in two locations on Earth on the North-South axis, you will notice the Sun has a different elevation above the horizon, which means different lengths of shadows will be cast on the ground. This is also a way to prove that the Earth is in fact round because a flat Earth would have equally-sized shadows everywhere at any given time of day.

## 2.2 Models of Earth

Here is a simple thought experiment to consider. Suppose you are trying to measure your own height. You probably have not given much thought about how to technically do this because it seems intuitive: place a measuring tape at the bottom of your feet and mark the measurement at the top of your head. If we break this down, there are some important rules to follow (Figure 2.2):

1. The measuring tape must originate somewhere. In other words, we need to define a reference point or surface of zero height (i.e., the ground).
2. The measuring tape must be a straight line and originate at a 90-degree angle, perpendicular to the ground.
3. The measurement must terminate at a point along an imaginary line that is tangential to your head, and yes, that line must be perpendicular to the measuring tape and also parallel with the ground.

Whenever you measure your height, the ground is easy to define. It is whatever point you are standing on. This starting point is also known as a **datum**. A datum is simply a reference point, set of points, or a surface from which distances can be measured. It does not matter if you are below sea level, atop Mount Everest, or on the 30th floor of a skyscraper. You will always get an accurate and repeatable measure of your height using a datum that is defined directly below your feet. But what about measuring the height of terrain on Earth? Whenever we measure the height of Earth's terrain above some reference surface, we are measuring **elevation**.

The same rules above apply when we measure elevation. In order for elevation measurements to be comparable across the world, we need to define a reference surface, a datum, for the entire planet. There are actually several ways that we can model the shape of Earth in order to produce a datum. Models of Earth's shape are often referred to as either vertical datums (the plural of datum) if you are referencing elevation or horizontal datums if you are referencing location. A **vertical datum** is a 3D surface model that is used to reference heights or elevations for the Earth. A simple question like *How high is Mount Logan in Yukon, Canada?* is complicated by the need for a reference surface and the fact that Earth's shape is irregular. In this section, we will review three types of

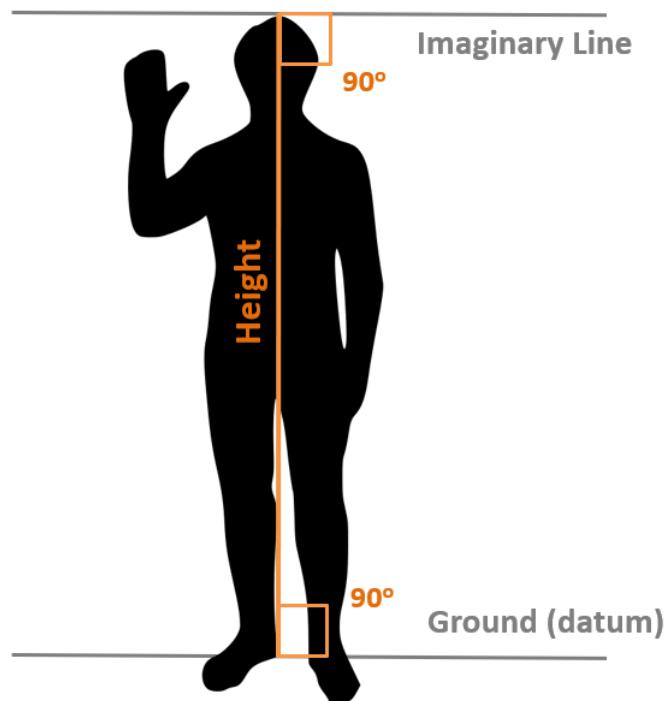


Figure 2.2: Diagram for measuring height above a datum. Pickell, CC-BY-SA 4.0.

vertical datums:

- Geodetic - based on geometry
- Tidal - based on sea level
- Gravimetric - based on gravity

### 2.2.1 Geodetic Vertical Datums

A **geodetic vertical datum** is one that describes the Earth's shape in the simplest possible terms using standard geometry. Despite what a globe might lead you to believe, the Earth is not perfectly spherical, but it is close to being spherical. In fact, the radius of Earth varies by no more than 22 km or 0.35%, hardly anything you would ever notice if you were holding it in your hand. That small difference is, however, significant enough to lead to mapping inaccuracies at the local level if a spherical model of Earth was adopted (Figure 2.3). Instead, we frequently describe Earth's shape as an oblate ellipsoid, which is essentially a sphere that has been flattened, and we define this ellipsoid with a semimajor and semiminor axis. Sometimes you will see the term *spheroid* used, which just means "sphere-like" and is interchangeable with the term *ellipsoid*.

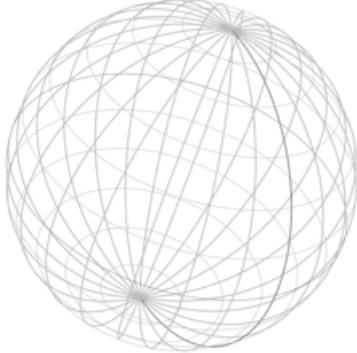


Figure 2.3: Spherical geodetic datum. Pickell, CC-BY-SA-4.0.

There are many different ellipsoids that have been defined and are currently in use as datums. The most commonly used ellipsoid is called the World Geodetic System of 1984 or usually abbreviated as WGS 1984 or WGS 84. In fact, there are hundreds of ellipsoids that have been defined over recent centuries to model the shape of the Earth. The reason for so many other ellipsoids is due in part to technological advances that have improved the accuracy and precision of surveying as well as estimation of the ellipsoidal parameters. Many of these ellipsoids are not **geocentric**, that is, not originating from the center of mass of Earth. These datums are known as **regional datums**, which still describe the dimensions that approximate the shape of Earth, but are instead oriented so that the surface of the ellipsoid is congruent with a particular regional surface

of Earth. For example, the European Datum 1950, the South American Datum 1969, the North American Datum 1983, and the Australian Geodetic Datum 1966 conform well to their respective continents, even better than WGS 1984 in most cases, but poorly anywhere else in the world.

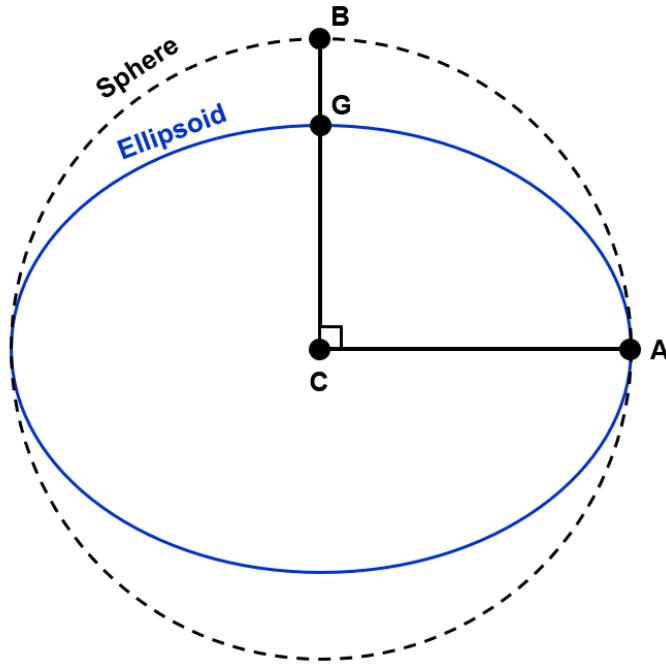


Figure 2.4: Sphere versus ellipsoid. Pickell, CC-BY-SA-4.0.

Figure 2.4 greatly exaggerates the flattening of the ellipsoid to illustrate the above points. In reality, the sphere is flattened using a flattening factor calculated as  $f = (CA - CG)/CA$  and defined exactly as  $f = 298.257223560$  for WGS 1984. Thus, the semiminor axis (i.e., rotational axis) for the WGS 1984 ellipsoid (meters) is

$$CG = CA - \left( CA \times \frac{1}{f} \right) = 6378137 - \left( 6378137 \times \frac{1}{298.257223560} \right) = 6356752.3$$

where  $G$  is the North Pole and  $A$  is a point on the Equator. The sphere, of course, is much simpler where  $\text{radius} = CB = CA = 6378137$ .

### 2.2.2 Tidal Vertical Datums

A **tidal vertical datum** is likely one that you are familiar with. The premise of a tidal vertical datum is to use mean sea level as a reference surface, above

which are positive elevations and below are negative elevations. This has a lot of advantages, like it is intuitive and oceans cover more than 70% of the planet's surface so much of Earth's land mass is near an ocean. However, the disadvantages are that sea level changes over time with tides and also with climate change. The not-so-obvious problem with a tidal vertical datum is that the sea level is actually not constant around the planet not only due to tides, but also temperature, air pressure, and gravity. In other words, mean sea level measured at a gauge station in Halifax on the Atlantic Ocean will not be the same distance from the center of Earth as mean sea level measured at Victoria on the Pacific Ocean (Figure 2.5). The primary challenge with a tidal vertical datum is extending it away from the coastline through a network of survey points using a process known as levelling, and even still, it is only meaningful during the epoch in which the mean sea level was measured at a number of tidal gauge stations.

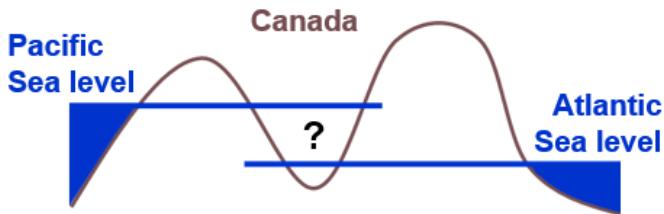


Figure 2.5: Conceptual tidal datum for Canada. Pickell, CC-BY-SA-4.0.

### 2.2.3 Gravimetric Vertical Datums

The **geoid** is a physical approximation of the figure of Earth. The shape represents Earth's surface with calmed oceans in the absence of other influences such as winds and tides. It is computed using gravity measurements of Earth's surface and is best thought of as the surface or shape that the oceans would take under the influence of Earth's gravity and rotation alone. In other words, the geoid represents the shape Earth would take if the oceans covered the entire planet. More specifically, the geoid is a **gravimetric** model of Earth's shape that is defined as an equipotential surface from a constant gravity potential value. Due to the distribution of mass on Earth, gravity is not constant across the planet's surface. As a result, the surface of Earth's oceans is not smooth like a sphere, but instead undulates depending on where gravity forces water to remain at rest. You can think of Earth's gravitational field as a series of parallel lines extending outwards from the center of mass of Earth into space. Any of these lines that you choose is an equipotential surface where the force of gravity is constant. Keep in mind that the force of gravity is stronger nearer the center of mass of Earth and weaker as you move away from it. Thus, the geoid is an arbitrary equipotential gravity surface that is chosen to roughly coincide with

present-day mean sea level.

When you measure the height of something relative to a gravimetric vertical datum like the geoid, you must level your instrument. Levelling forms a vertical line that is orthogonal or perpendicular to the geoid, known as a **plumb line**. It is incredibly easy to visualize a plumb line. Simply tie a rock to the end of a string and hold the string with your outstretched arm. The length of the straightened string traces a plumb line to the center of mass of Earth, wherever you are. Because gravity changes with location on Earth and all plumb lines are converging on a singular point, plumb lines are never parallel. This phenomenon has important implications for comparing observations on the ground with a geodetic model of Earth like an ellipsoid. In other words, the plumb line that you traced with your string is pointing to the center of mass of the geoid, but the center of the ellipsoid is often in a slightly different direction. This difference is known as the **deflection of the vertical** and is measured as the angular difference between the centre of the geoid and the centre of a reference ellipsoid. Like other measurements of geodetic location (i.e., latitude and longitude), the deflection of the vertical is comprised of two angles:  $\xi$  (xi) representing the north-south angular difference and  $\eta$  (eta) representing the east-west angular difference.

It should be evident by now that the reference surface that you choose as a vertical datum will determine the measured elevation of Earth's terrain. Additionally, we frequently need to convert elevations between geodetic and gravimetric vertical datums. For example, when you use a Global Navigation Satellite System receiver, you are provided with an elevation that is relative to the WGS 1984 ellipsoid (more on that in Chapter 4). The difference in height between an ellipsoid and the geoid is referred to as **geoid height (N)** while the difference in height between an ellipsoid and Earth's surface is referred to as **geodetic or ellipsoidal height (h)**. The difference in height between the geoid and the Earth's surface is called **orthometric height (H)** (Figure 2.6), and is given as:

$$H = h - N$$

To illustrate the concept of a gravimetric datum, suppose we constructed a large, straight tunnel through the physical Earth that was tangential to the ellipsoid. If we allowed the oceans to flow freely through this tunnel, your experiences might convince you that water would flow from one end to the other. But in fact, this tunnel is so large, that the gravity field is changing. So the water would actually come to rest at the surface of the geoid or gravimetric model, as shown in Figure 2.7 below.

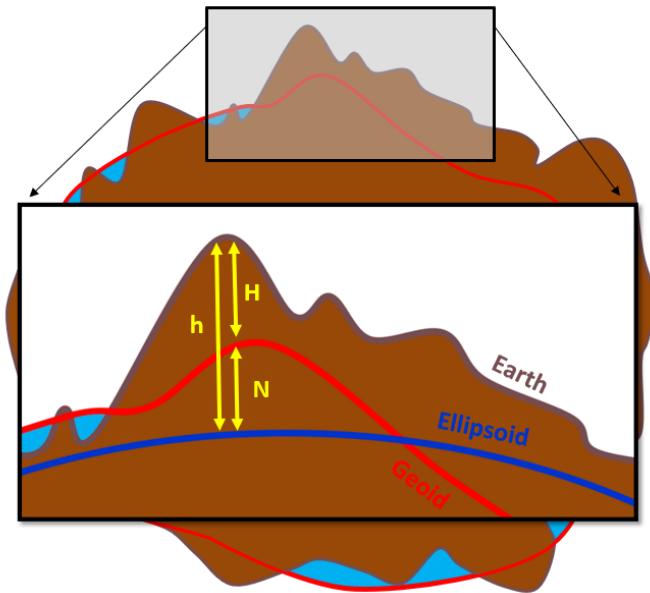


Figure 2.6: Orthometric Height ( $H$ ) is the ellipsoidal height ( $h$ ) less the geoid height ( $N$ ). Pickell, CC-BY-SA-4.0.

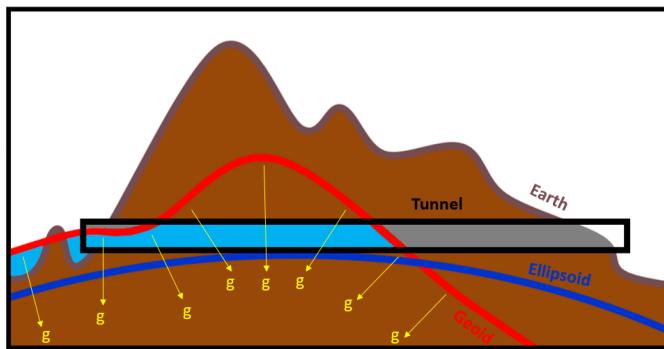


Figure 2.7: Thought experiment showing where water would be at rest within a tunnel through the geoid due to the equipotential force of gravity ( $g$ ). Pickell, CC-BY-SA-4.0.

## 2.3 Case Study: The Canadian Geodetic Vertical Datum of 2013

The Canadian Geodetic Vertical Datum of 2013 (CGVD2013) is the current gravimetric vertical datum used in Canada to reference heights. It is defined with a potential gravity value of  $62,636,856.0 \text{ m}^2 \text{ s}^{-2}$ . The previous vertical datum in Canada - the Canadian Geodetic Vertical Datum of 1928 (CGVD28) - was actually a tidal vertical datum that corresponded to mean sea level measured at Yarmouth, Halifax, Pointe-au-Père, Vancouver and Prince-Rupert, and a height in Rouses Point in New York. It turns out that Halifax referenced to CGVD2013 is 64 centimeters *below* Halifax referenced to CGVD28!

For reference, CGVD2013 is 17 centimeters below mean sea level measured in Vancouver at the Pacific Ocean, 39 centimeters above mean sea level in Halifax at the Atlantic Ocean, and 36 centimeters above mean sea level in Tuktoyaktuk at the Arctic Ocean. The older CGVD28 did not have any survey benchmarks in the far north of Canada and, with the advent of more reliable satellite-based measurements, was modernized in 2015 to CGVD2013. The United States currently uses the North American Vertical Datum of 1988 (NAVD88), which was never adopted by Canada, but the United States will be modernizing their vertical datum by adopting a gravimetric model with the same gravity potential value as Canada as early as 2025.

## 2.4 Referencing Location

### 2.4.1 Cartesian Coordinate Systems

Now that we have explored how to reference heights to vertical datums, we will now turn to considering how to reference location on Earth. Before we jump to the three dimensional case of Earth, consider how you would map your room and identify your location within that room. Assuming you are in a rectangular room, you could easily pick a corner and first measure the distances between the four corners of the room, giving you the dimensions. You could then proceed to measure your distance to any two walls and quite easily define your position within the room relative to the first corner that you picked. This is an example of a **coordinate system** that provides reference for the relative locations of anything contained within the extent of the coordinate system (i.e., the four corners of your room).

Fundamentally, a coordinate system is defined by a common unit of measurement (e.g., meters, feet, degrees), an orientation defining the direction that measurements positive or negative, and an **origin**, which is an arbitrary point where the measurements begin at zero. On a one dimensional line, you can define any other location on the line as a measured distance from the origin at [0]. On two dimensional maps, like your room, we have two axes that are perpendicular from which we can define any location with measured distances from

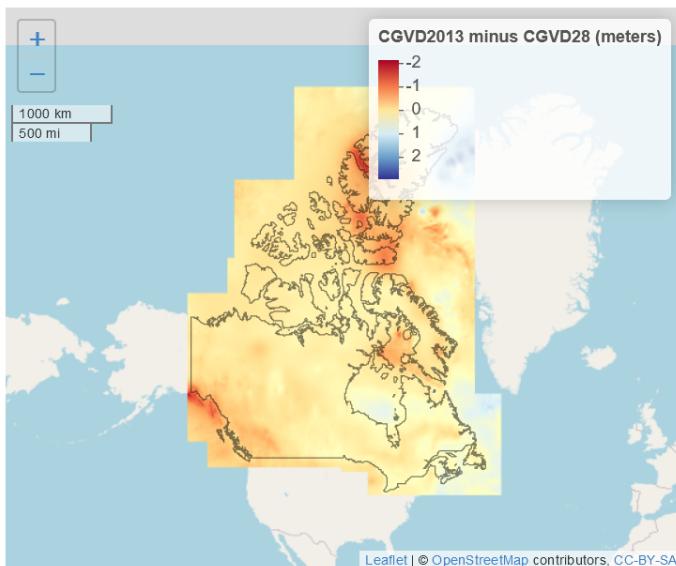


Figure 2.8: The difference (in meters) between the Canadian Geodetic Vertical Datum of 2013 (CGVD2013) and the Canadian Geodetic Vertical Datum of 1928 (CGVD28). Pickell, CC-BY-SA-4.0. Data from Natural Resources Canada and licensed under the Open Government Licence - Canada. Animated figure can be viewed in the web browser version of the textbook: <https://ubc-geomatics-textbook.github.io/geomatics-textbook/mapping-data.html>

the origin  $[0, 0]$ . We can also extend this to the three dimensional Earth, which simply requires another axis to define the origin at  $[0, 0, 0]$ . All of these cases are referred to as **Cartesian coordinate systems**, so-named after French philosopher Descartes (also known for the phrase, “I think, therefore I am”) who first described the two dimensional case in 1637 in *La Géometrie*.

### 2.4.2 Celestial Coordinate Systems

You might be wondering, *How do we reference locations on Earth?* Before the technological era of geoids, astronomical observations provided the basis for the earliest coordinate systems. The ancient civilizations of Greece, Egypt, and Babylon all recognized the use of celestial coordinate systems for defining the location of stars, planets, and other celestial bodies in the sky. Recall that Eratosthenes, a Greek mathematician living about 2300 years ago, had already worked out the spherical shape of Earth. A celestial coordinate system simply extends the spherical shape of the Earth outward into space to locate objects using angular measurements. It is a *geocentric* coordinate system, that is based on an origin at the center of Earth with an orientation following the rotational axis of Earth (i.e., spinning around the semiminor, North-South axis). This is also known as an equatorial coordinate system, because it is oriented relative to the Equator of Earth. The equatorial coordinate system is the reason why you can navigate by Polaris, the North Star, which would appear nearly directly overhead if you were standing at the North Pole. Even though Earth is not perfectly spherical, astronomical observations have been reliably used for millennia to transit the irregular oceans and terrain of Earth.

### 2.4.3 Geographic Coordinate Systems

Geocentric coordinate systems are essential for global navigation and mapping. In the modern era, we use **geographic coordinate systems** (sometimes abbreviated GCS) that are oriented to the rotational axis of Earth, much like the equatorial coordinate system. The origin of a geographic coordinate system is the center of Earth and the units of measurement are degrees of **longitude** and degrees of **latitude**. Degrees of latitude (denoted by lambda,  $\lambda$ ) measure the angle from the equitorial plane North (+) or South (-), while degrees of longitude (denoted by phi,  $\varphi$ ) measure the angle from the polar plane East (+) or West (-). Thus, geographic coordinate systems use angular units of measurement. Any combination of latitude and longitude gives coordinates  $[\lambda, \varphi]$  on a sphere or ellipsoid. Positive values of latitude put you in the Northern Hemisphere, while negative values of longitude put you in the Western Hemisphere. Constant lines of latitude, known as **parallels** because they are always parallel to one another, and lines of longitude, known as **meridians**, form a grid that fits over a sphere or ellipsoid called a **graticule**.

For most of your life, you have probably believed that there is a singular combination of capital-L Latitude and Longitude values that absolutely define some

location on Earth in perpetuity. This is perhaps the most profound geographic lie that we were taught as young school children. In fact, there are as many “types” of latitude as there are geographic coordinate systems. By now, you should be able to recognize the difference between *geocentric latitude* that is referenced to a sphere and *geodetic latitude* that is referenced to an ellipsoid. The main difference is that geocentric latitude is the angle relative to the centre of the sphere at the equatorial plane and geodetic latitude is the angle relative to the equatorial plane (i.e., not necessarily the centre). Figure 2.9 illustrates how the same angle can put you in two very different places depending on the geographic coordinate system you are using.

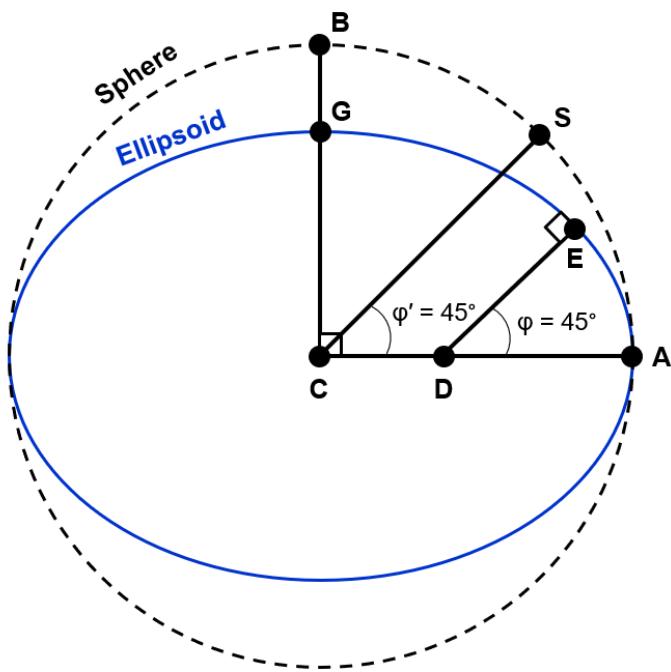


Figure 2.9: Geocentric versus geodetic latitude. Pickell, CC-BY-SA-4.0.

$CS$  represents the line connecting the centre of a spherical geographic coordinate system to the surface of the sphere at a geocentric latitude  $\varphi' = 45^\circ$  and  $DE$  represents the line connecting the equatorial plane to the surface of the ellipsoid at a geodetic latitude  $\varphi = 45^\circ$ . Notably,  $CS$  is parallel to  $DE$  and  $CS$  intersects with the surface of the sphere at a  $90^\circ$  angle and  $DE$  intersects the surface of the ellipsoid at a  $90^\circ$  angle.

A coordinate system that is referenced to an ellipsoid is known as a **horizontal datum**. For example, the World Geodetic System of 1984 (WGS 1984) is a *geodetic* datum that has a geographic coordinate system referenced to it. So the origin of WGS 1984 is also the origin of the GCS from which latitude and longitude measurements are derived. So if you measured your latitude and longitude

using the WGS 1984 horizontal datum, you would be placed somewhere on the ellipsoid also defined by WGS 1984. It is important to note, however, that the same exact *geocentric latitude* and longitude measures would place you somewhere else on a sphere. Other factors such as plate tectonics, glaciation, and ocean tides cause the Earth's surface to be in constant motion underneath any fixed horizontal datum. For example, Europe has drifted about 60 meters from North America since Eratosthenes first calculated the circumference of Earth. Therefore, any "type" of latitude or longitude is only useful during a particular epoch of time.

If you mapped the Earth as an ellipsoid in a three dimensional Cartesian coordinate system, you could describe location using three coordinates  $[x, y, z]$ , with  $[0, 0, 0]$  being the center of Earth. However, we do not often express mapped coordinates of Earth using Cartesian coordinates. Instead of referring to the North Pole at  $[0, 0, 6356752.3]$  meters (the polar radius of Earth) we usually refer to it with a single coordinate,  $90^\circ$  N. Why? The North and South Poles are the only points on Earth that can be defined with a single coordinate because they coincide with the orientation of a geographic coordinate system. Imagine placing a protractor at a  $90^\circ$  angle relative to a table. If you rotate it along that perpendicular axis, one arm of the protector would spin in a  $360^\circ$  circle, but the other arm of the protractor would always point up or down at  $90^\circ$  in the same direction. For all other locations on Earth, a pair of  $[\lambda, \varphi]$  coordinates are needed to define location. With space-based global navigation systems it is more common to combine coordinates in both horizontal and vertical datums together. For example,  $[\lambda, \varphi]$  expresses your location relative to the horizontal datum while  $[\lambda, \varphi, h]$  expresses your ellipsoidal height ( $h$ ) at a location and  $[\lambda, \varphi, H]$  expresses your orthometric height ( $H$ ) at a location.

### Your Turn!

Use our geodesy tool to visualize the differences between the sphere, ellipsoid, and geoid. You can modify the transparency of the sphere and ellipsoid to see how these geometries vary. Change the flattening of the ellipsoid to achieve different models of Earth. Search for locations with latitude and longitude values and calculate the geoid height. Compare geocentric and geodetic latitudes. The sphere and the ellipsoid models in the visualization represent Earth's equatorial diameter (semi-major axis) as 6,356,752.3 m (i.e., equivalent to WGS 1984) and the geoid elevations are exaggerated by a factor of 0.549 to enhance their appearance.

#### 2.4.4 Projected Coordinate Systems

Despite everything covered so far in this chapter, we very rarely see or display geographic data in 3-dimensions. In fact, most geographic data you will likely encounter will be either 1- or 2-dimensional (more on that in Chapter 3). Geographic coordinate systems are incredibly important for understand-

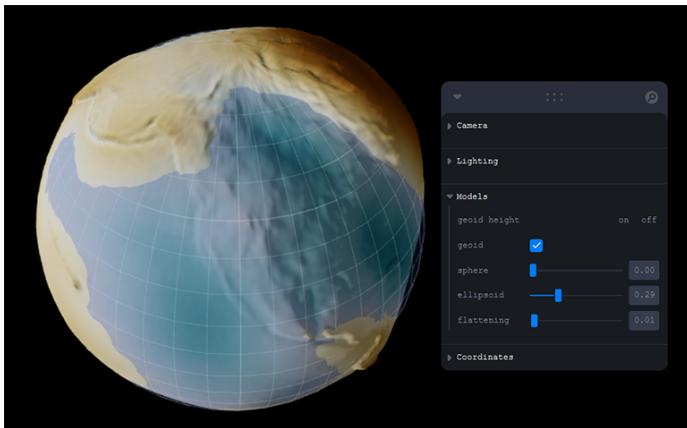


Figure 2.10: Online geodesy tool for visualizing the differences between the sphere, ellipsoid, and geoid. <<http://oergeomatics-geodesy-viz.s3-website.ca-central-1.amazonaws.com/>> Click here to access the interactive tool.</a> Floria Gu, CC-BY-SA-4.0.

ing how geographic data are fundamentally “attached” to the Earth. However, geographic coordinate systems are not suitable for creating maps, those 2-dimensional spatial models that can be easily displayed on your computer screen or printed on a sheet of paper. Instead, cartographers rely on **projected coordinate systems** that flatten a 3-dimensional geographic coordinate system to a 2-dimensional map. Really, these projected coordinate systems involve transformations called **map projections** that convert 3-dimensional coordinate space into 2-dimensional coordinate space, which means the map units are linear such as meters. Whenever we move from a geographic to a projected coordinate system, we lose information, and distortion results.

Cartographers have wrestled with how to project Earth onto a printed page for millennia. The fundamental mathematics for map projections were first comprehensively described by Claudius Ptolemy around 150 C.E. Ptolemy’s work *Geography* was one of the earliest treatises on cartography and map making that included an atlas of regional maps of Europe, Africa, and Asia. Ptolemy’s work built on and came several centuries following Eratosthenes and earlier Greek geocentric observations by Plato and Aristotle. Ptolemy observed that a globe was the best way to represent the intervals and proportions of Earth’s surface without distortion. However, globes are not very useful for looking at regions in detail and you can only see part of a globe at any given time. Thus, a mathematical language is needed to translate a geographic coordinate system to a planar or projected coordinate system. *Geography* was lost to antiquity before it was rediscovered, copied, and translated centuries later, first by Muslim cartographers in the 9th century C.E. and later by Italian cartographers in the 15th century C.E. during the Renaissance, which gave rise to the many types of

map projections that we see today.

Because all map projections result in distortion from the loss of the third spatial dimension, it is useful to think about map projections in terms of what they preserve. There are four main characteristics that can be distorted or preserved, which give rise to the primary types of map projections that are in use for environmental management applications:

- **Conformal** projections preserve shape and angles
- **Equal-area or equivalent** projections preserve area
- **Azimuthal** projections preserve direction
- **Equidistant** projections preserve scale and distances

Some map projections can preserve several of these characteristics at once, but only a globe can simultaneously preserve area, direction, distance, and shape. Any map projection will have inherent trade-offs representing these characteristics accurately. It is beyond the scope of this textbook to discuss all map projections. Instead, we will focus on several key examples of map projections that are commonly used for environmental management applications. For a more comprehensive discussion of map projections generally, the reader is referred to *Map Projections: A Working Manual* by [Snyder, 1987]. In the next section, we look at how map projection distortion can be measured.

#### 2.4.5 Measuring Map Projection Distortion

**Tissot's Indicatrix** is often used to visualize distortion from map projections, named after Nicholas Auguste Tissot. The metric is relatively simple: Tissot's Indicatrix is a perfect circle on the surface of a 3-dimensional globe, but will form an ellipse whenever projected to a 2-dimensional coordinate system. For this reason, Tissot's Indicatrix is sometimes referred to as Tissot's Ellipse. Since ellipses can vary along two axes, Tissot's Indicatrix can represent areal, angular, and linear map distortions both longitudinally and latitudinally at any location in the map. This is very handy, because we can place Tissot's Indicatrices (the plural of indicatrix) at different locations and examine how distortion changes throughout the map projection.

So how do we use this tool? The quotient between a line projected onto a map  $a$  and the same line on a globe  $a'$  is  $\frac{a}{a'} = 1$  when there is no distortion on that axis of the ellipse. This quotient is also called a scale factor because it is showing how the map scale is modified locally by a map projection. Figure 2.11 below shows an example of a reference indicatrix that is a perfect circle on a globe with axes  $a$  and  $b$ .

Let us assume that  $a = b = 1$ . Then, the reference indicatrix has the following properties:  $a = b$ ,  $a \times b = 1$ , and  $\text{Area} = \pi ab = \pi^2$ . If  $\frac{a'}{a} > 1$ , then we can conclude that the projection is *expanding* the distance along the  $a$  line. If  $\frac{a'}{a} < 1$ , then we can conclude that the projection is *compressing* the distance along the line  $a$ . For example, suppose  $a = 2$  and  $b = 0.5$ , then we have modified

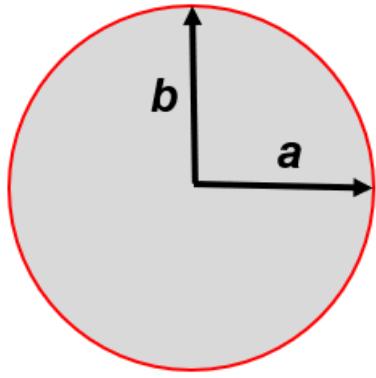


Figure 2.11: Reference Tissot Indicatrix. Pickell, CC-BY-SA-4.0.

the scale of the indicatrix along both axes, but the areal solution is the same as the reference indicatrix shown as the red dashed circle in Figure 2.12 below.

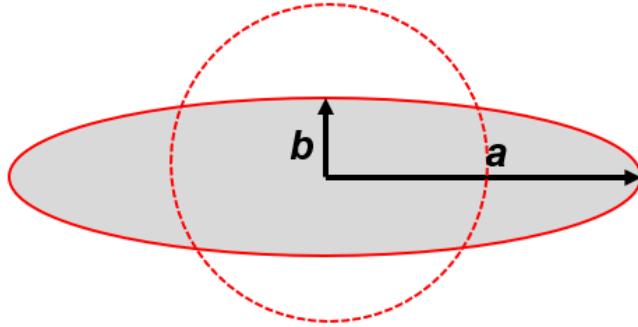


Figure 2.12: Equivalent Tissot Indicatrix. Pickell, CC-BY-SA-4.0.

The indicatrix above in Figure 2.12 is an example of an *equivalent* indicatrix, which has the following properties:  $a > b$ ,  $a \times b = 1$ , and  $\text{Area} = \pi ab = \pi^2$ .

If  $\frac{a}{a'} \times \frac{b}{b'} > 1$ , then we can conclude that the projection is *inflating* the area. If  $\frac{a}{a'} \times \frac{b}{b'} < 1$ , then we can conclude that the projection is *deflating* the area. Consequently, whenever the quotients of both axes are equivalent (i.e.,  $\frac{a}{a'} = \frac{b}{b'}$ ), Tissot's Indicatrix forms a perfect circle and the ellipse is conformal with angles true to the globe. For example, suppose  $a = b = 2$ , then we have modified the scale of the indicatrix along both axes, but with the same factor. This results in a *conformal* indicatrix that is not equivalent, shown in Figure 2.13 below.

This conformal indicatrix has the following properties:  $a = b$ , but  $a \times b \neq 1$ , and therefore  $\text{Area} = \pi ab = 4\pi^2$ . As you can see, whenever  $a \times b = 1$  the indicatrix is equivalent (equal-area) and whenever  $a = b$  the indicatrix is conformal.

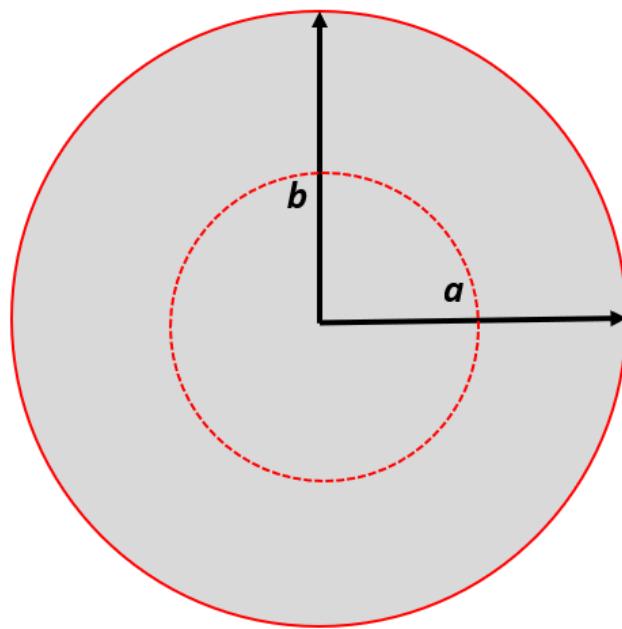


Figure 2.13: Conformal Tissot Indicatrix. Pickell, CC-BY-SA-4.0.

## 2.5 Map Projections for Environmental Management

### 2.5.1 Mercator

Mercator is a cylindrical map projection that represents meridians and parallels as straight lines. The cylindrical surface is oriented such that the rotational axis of Earth runs through the openings of the cylinder and the Equator represents the tangent where the cylinder meets the Earth's surface. Scale along the tangent is true because the translation from the spherical Earth to the cylindrical surface is one-to-one at the tangent, so this is also the location on the projection where there is no distortion. This has the effect of accurately representing the shape and angles on the map (i.e., conformal), but greatly distorts area as you move away from the Equator. In fact, the North and South Poles are represented as 2-dimensional lines at the top and bottom edges of the map instead of 1-dimensional points. Although scale and area change as you move North or South along a meridian, scale and area are equivalent along any parallel, but not necessarily true to a globe. In other words, you can compare area or scale anywhere along a parallel, but only at the Equator is the area and scale true to the globe.

The Mercator map projection is perhaps the most pervasive and reproduced pro-

jection around us (Figure 2.14). Because angles are preserved, you can easily and accurately navigate long distances across Earth, and this was exactly the purpose that Gerardus Mercator envisioned when he first identified the projection for sea-faring Europeans in 1569. You may also recognize the Mercator projection from web mapping applications like Google Maps, which use it because it ensures that North-South roads intersect at right angles with East-West roads.

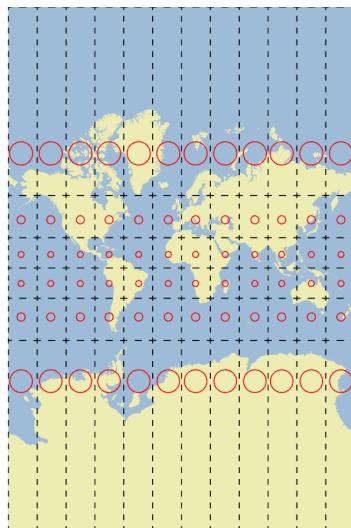


Figure 2.14: Mercator map projection with Tissot’s Indicatrices in red. Pickell, CC-BY-SA-4.0.

### 2.5.2 Universal Transverse Mercator (UTM)

Universal Transverse Mercator (UTM) is very similar to a Mercator projection except that the cylinder is rotated or transverse by  $90^\circ$  so that the opening of the cylinder is perpendicular to the rotational axis of Earth. This has the effect of moving the tangent from the Equator to any Meridian. In fact, you can rotate the cylinder at any angle you want where  $0^\circ$  is a true Mercator,  $90^\circ$  is a transverse Mercator, and any other angle is considered an oblique Mercator. UTM is actually a system of 60 different transverse Mercator projections that are defined to represent  $6^\circ$  Longitudinal intervals of Earth’s surface ( $60 \text{ zones} \times 6^\circ = 360^\circ$ ). Each projection is defined as a zone, which is also divided into North and South zones depending whether you are in the Northern or Southern Hemisphere. Canada spans 15 UTM zones from Zone 7 North in the Yukon to Zone 22 North covering Newfoundland. Figure 2.15 below shows a map of UTM Zone 13 over Saskatchewan.

Besides defining the orientation of the cylinder, we can also specify its size or diameter. When the diameter of the cylinder is equivalent to the diameter of

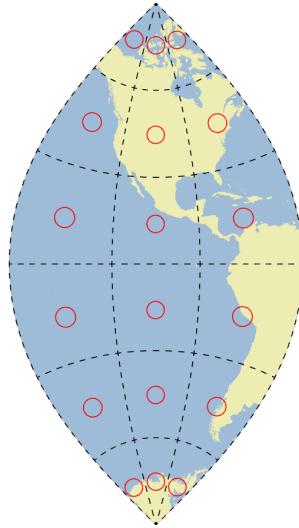


Figure 2.15: Universal Transverse Mercator Zone 13 map projection with Tissot’s Indicatrices in red. Pickell, CC-BY-SA-4.0.

Earth at the Equator, a single tangent line is formed. If the diameter of the cylinder is smaller than Earth’s diameter at the Equator, then the cylinder has two lines that contact Earth’s surface, known as **secants**. The purpose for having two secants is that the projection distortion can be more evenly distributed across the map. In the case of UTM, a scale factor of 0.9996 is applied to shrink the transverse cylinder slightly, forming two secants that are 360 km apart East-West. In between the two secants lies the **central meridian**, which is used to define the origin for the projected coordinate system. It is important to realize that the secants are parallel to each other and the central meridian, which means the secants are not meridians on Earth and form what are called **small circles**, a line that does not divide Earth into two equal portions.

UTM uses a unique coordinate system that deserves some explanation. Just like with latitude and longitude coordinates  $[\lambda, \varphi]$ , an arbitrary origin needs to be defined so that we know where we are in relation to the origin on the map. For projected coordinate systems like UTM that use linear units of measure such as meters, the origin is defined as the intersection of the central meridian and the Equator. Simple enough, right? There is one catch: UTM does not use any negative coordinates by convention. Thus, the origin of the coordinate system for each zone must be moved so that coordinates West of the central meridian and South of the Equator are positive. To do this, a constant value is added to all East-West coordinates (known as **Eastings**) and all North-South coordinates (known as **Northings**) to create what are known as **False Eastings** and **False Northings**, respectively. A value of 500,000 m is added to all Eastings so that the western limit of the zone is located at 0 m, the central meridian is located at

500,000 m, and the eastern limit of the zone is located at 1,000,000 m. A value of 10,000,000 m is added to all Northings in the Southern Hemisphere so that the Equator is at 10,000,000 m and 0 m is near the South Pole for all southern UTM zones. You might recognize the importance of the 10,000,000 m value because this represents approximately one-quarter of the Earth's North-South circumference.

### 2.5.3 Sinusoidal

Sinusoidal is a *pseudocylindrical* map projection, so-named because these projections approximate a true cylindrical projection except that Meridians are curved instead of straight like with Mercator or UTM. Because Meridians are curved, Sinusoidal maps represent the North and South Poles as single points instead of lines as is the case with Mercator. Thus, Sinusoidal maps are not conformal and distort shape, but they are in fact equal-area (Figure 2.16). Equal-area map projections like Sinusoidal are important for accurately accounting for land cover and other global mapping efforts. Therefore, it is common to find global datasets distributed in a Sinusoidal map projection.

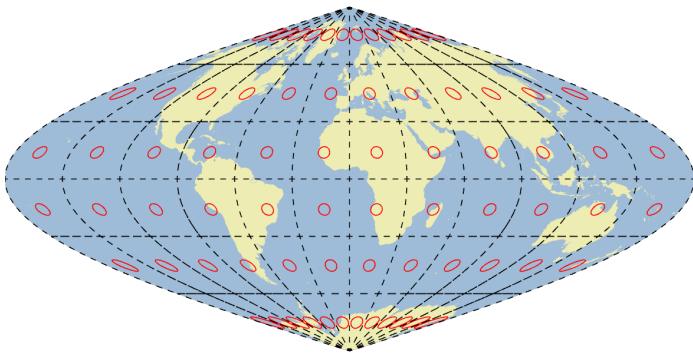


Figure 2.16: Sinusoidal map projection with Tissot's Indicatrices in red. Pickell, CC-BY-SA-4.0.

### 2.5.4 Albers

Albers is another example of an equivalent map projection, but unlike Sinusoidal, an Albers projection uses a cone as a projection surface instead of a pseudocylinder. Like the cylindrical case, the cone can be sized and oriented in any way, but the cone of an Albers projection is typically oriented so that

the vertex of the cone aligns with the rotational axis of Earth. The base of the cone has a diameter that usually results in two secants, known as **standard parallels** on an Albers map. Thus, Albers projections tend to distort latitudinally as you move North-South away from the standard parallels, but even more so as you move toward the base of the cone (Figure 2.17). Besides being an equal-area projection, Albers is a good choice for mapping regions because shape and scale are mostly preserved near the standard parallels. For example, the province of British Columbia has adopted a modified Albers projection that situates the standard parallels at  $50^{\circ}$  N and  $58.5^{\circ}$  N, which are near the northern and southern latitudinal limits of the province. This narrow band of latitude between the standard parallels ensures that there is relatively little distortion in shape and scale within the province, which is comparable to UTM. However, British Columbia is a longitudinally wide province, spanning 6 of Canada's 15 total UTM zones, so Albers has a distinct advantage of being able to show the entire province with little distortion. For the same reasons, you will often find Canada-wide data distributed in an Albers projection.

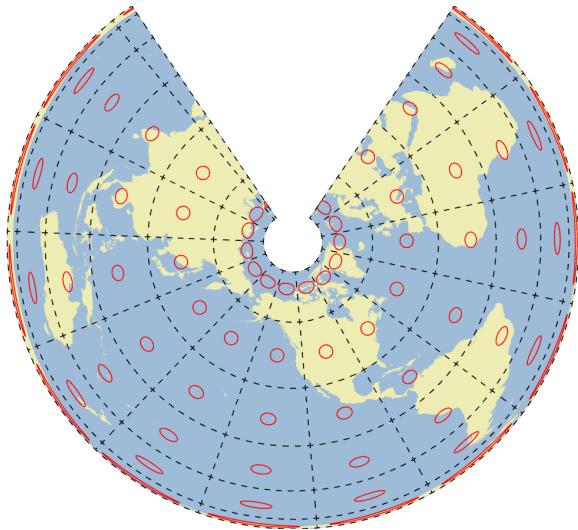


Figure 2.17: British Columbia Environment Albers map projection with Tissot's Indicatrices in red. Pickell, CC-BY-SA-4.0.

### 2.5.5 Azimuthal

Azimuthal projections use a flat circular plane to project the Earth onto a map. This plane is usually oriented so that it is tangent at a single point, usually the North or South Pole. In the case of a polar azimuthal projection, meridians radiate outward as straight lines from the pole to the edge of the circular plane and the parallels are represented as concentric circles. As a result, distortion increases as you move away from the centre point of the map with the outer

edge of the plane representing an **antipode** or an opposite point on Earth. The primary benefit of azimuthal projections is that they preserve direction and distance between the centre point and any other point in the map (Figure 2.18). The shortest geographic distance between the centre point and any other point creates a line known as a **great circle**, which divides the Earth into two equal portions. So another benefit of an Azimuthal project is that great circles can be mapped as straight lines. Azimuthal projections are commonly used when distance and direction are important, such as weather RADAR stations or air traffic control towers. It is important to realize that the centre point for an azimuthal projection can be any point on Earth and the equidistant property can be exploited for a number of applications.

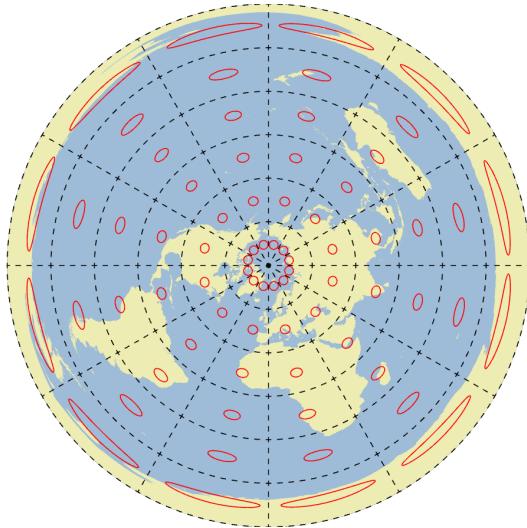


Figure 2.18: North Pole Azimuthal map projection with Tissot’s Indicatrices in red. Pickell, CC-BY-SA-4.0.

## 2.6 Summary

In this chapter, you learned about the science and technology of geodesy that goes into mapping data. We described the different models of Earth’s shape and the advantages and disadvantages that each model has. More generally, models of Earth represent vertical datums to which heights are referenced. When we think of “where” something is on Earth, we must use horizontal datums to reference location. There are two types of horizontal datums that are very important to geomatics: (1) a geographic coordinate system uses lines of latitude and longitude to define locations on Earth’s 3-dimensional surface while (2) a projected coordinate system flattens Earth into a manageable 2-dimensional coordinate space. In the case of projected coordinate systems, we have many

choices when deciding which map projection to use when we map data, each with its own uses and distortions. The next time that you look at a map, ask yourself with your new found appreciation of geodesy, how is this information being misrepresented to me?

### **Reflection Questions**

1. Describe the process of measuring the height of something on Earth.
2. Explain the difference between a geographic coordinate system and a projected coordinate system.
3. Name as many projected coordinate systems as you can.

### **Practice Questions**

## Chapter 3

# Data Types and Spatial Data Models

Written by June Skeeter and Paul Pickell

In the previous chapter, we discussed some of the unique challenges associated with representing spatial data in a GIS, and how to account for these with geographic coordinate systems and map projections. In this chapter we will discuss more broadly how to represent both spatial and non-spatial data in a Geographic Information System. We will introduce the different types of data that can represent non-spatial attributes and discuss the different scales this data can be measured on. Then we will introduce the different *spatial data models* we use to link the spatial and non-spatial data. Finally, we will cover some of the different file types that can be used to store data.

### Learning Objectives

1. Types of Spatial Phenomena
2. Measurement Scales of Data: Quantitative vs. Qualitative
3. Overview of Raster and Vector Data Models
4. Data Resolutions
5. Common File Types in GIS

### Key Terms

Phenomena, Discrete Object, Continuous Field, Qualitative, Quantitative, Measurement Scale, Raster, Vector, Resolution,

## 3.1 Types of Phenomena

**Phenomenon**, noun, plural *Phenomena*: 1 a fact or situation that is observed to exist or happen, especially one whose cause or explanation is in question. 2 a remarkable person, thing, or event [Oxford Languages]. Essentially, anything and everything are phenomenon: lightning, a country, coastlines, a dog on a kayak. Broadly speaking, in GIS we categorize phenomena as **discrete** or **continuous**. Both kinds of phenomena can be represented in a GIS, but they come with different considerations and cannot always be represented with the same kind of data model.



Figure 3.1: Yarrow enjoying the scenery at Alouette Lake, she's quite the phenomenon indeed. Skeeter, CC-BY-SA-4.0.

### 3.1.1 Discrete Objects

Discrete objects are finite and have distinct boundaries. Each object is a unique, self contained entity whose geography can be exactly defined. Because each object is unique and self contained, collections of objects are countable. A concrete example of a discrete objects would be buildings. They are real physical objects with well defined boundaries. We can count the number of buildings on a college campus or in a city. National and sub-national boundaries are also discrete objects. They (typically) have well defined boundaries and we can easily count the number of nations or provinces. They are not, however, real physical objects. Political boundaries are arbitrary human constructs.

### 3.1.2 Continuous Fields

Continuous fields are infinite and lack defined boundaries. Fields can be measured at an infinite number of locations. However, similar values tend to cluster in space so we can often make assumptions based on finite observations of continuous fields. One of the most common examples of a continuous field is elevation. This is a physical property associated with every location on earth.



Figure 3.2: Provinces are clearly delineated, distinct objects, despite having no real physical presence. Skeeter, CC-BY-SA-4.0.

We can't count the "number of elevations" because space is infinitely divisible and everywhere in space can have an elevation.

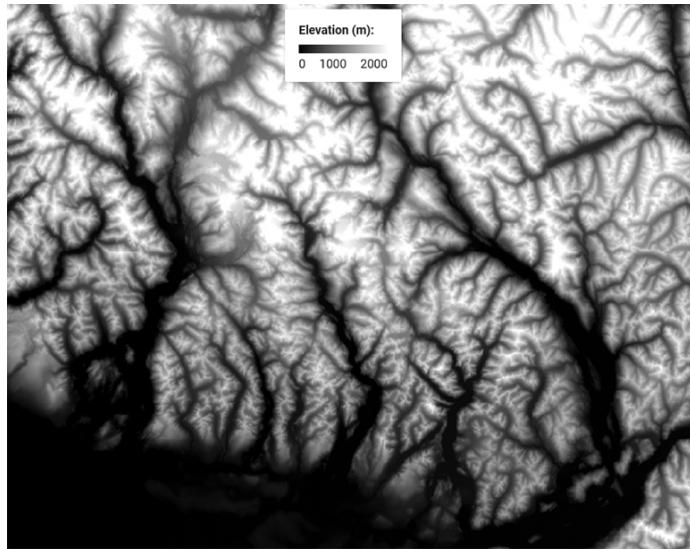


Figure 3.3: Elevation of the Sunshine Coast in BC, from the NRCAN DEM (google earth engine) [NRCAN, 2021].

### 3.1.3 Imperfect Distinctions

Few phenomena will fit perfectly and exclusively into one category or the other. That said, it's helpful for us to think about the discrete v. continuous dichotomy. As long as we recognize that it's not a perfect classification. Whether a phenomenon is considered discrete or continuous depends on scale (both spatial and temporal) and perspective. Some phenomena are a bit of both. Take the coastlines, they can be treated as discrete or continuous. At the scale of an individual beach over hours, tides can cause wide variations in water levels/position. How/where does one draw the discrete line representing the "coast". At this scale, the coast isn't really a discrete object, rather a continuous field known as the inter-tidal zone. Zoom out a bit and those fluctuations aren't particularly relevant if you want to make a map of Pacific Rim National Park. The coast could be considered a discrete object. But if you change the timescale and look at sea level rise projections, then you're dealing with a continuous field.

A lightning strike is an electric discharge between the atmosphere and the ground. A lightning strike is a discrete object. The precise location of the strike can be pin pointed, the number of strikes during a storm can be counted. But what about the actual lighting bolt? That's more a continuous field, it's not really possible to measure the exact boundaries of the path the electric discharge takes. Then we can look at other things, like the probability of lightning strikes.

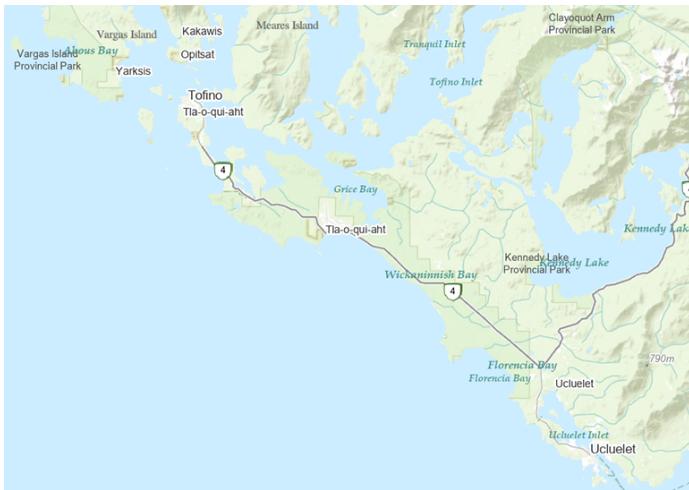


Figure 3.4: The west coast of Vancouver Island (just a screen grab from Arc, might want something better. Skeeter, CC-BY-SA-4.0.

Figure 3.5 is a continuous field, calculated from counting discrete objects.

## 3.2 Types of Data

Within the context of a Geographic Information System, each piece of information pertaining to a phenomena can be referred to as an **Attribute**. An phenomena can have many different attributes associated with it, but each attribute can broadly be said to address one of three questions: **What**, **When**, or **Where**? Attributes that describe *where* are known as **Spatial Data** while all other attributes are **Non-Spatial Data**. All data, spatial and non-spatial, can broadly be classified as either **qualitative** or **quantitative**. These data types are fundamentally different and are therefore measured on fundamentally different scales. The types of analysis we can conduct with qualitative data are more limited than quantitative data, but that does not necessarily mean quantitative data are “better” than qualitative.

### 3.2.1 Qualitative Data

Qualitative data are categorical; they are strictly descriptive and lack any meaningful numeric value. They describe the qualities of a phenomenon, without giving us any numeric information. Most qualitative data you will work with in GIS are textual or coded numerals, but there are circumstances where you may encounter non-textual data (e.g. images, sound clips, videos) in a dataset. Qualitative data can be “spatial” in nature (e.g. relative directional descriptors: left/right, near/far, north/south), but because they lack numeric values, they

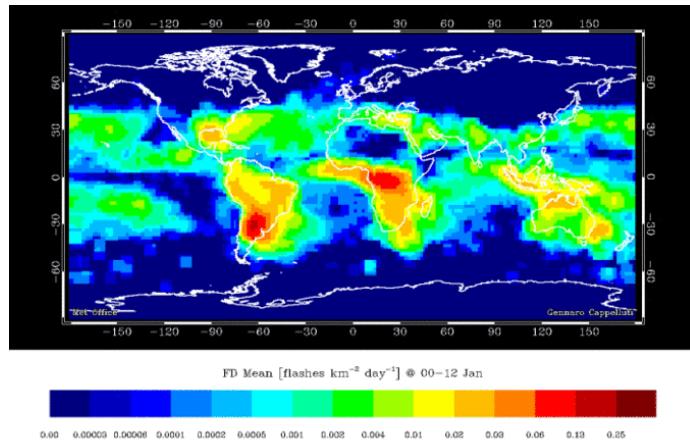


Figure 3.5: Global lighting strike density per month. Skeeter, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://ubc-geomatics-textbook.github.io/geomatics-textbook/types-of-data.html>

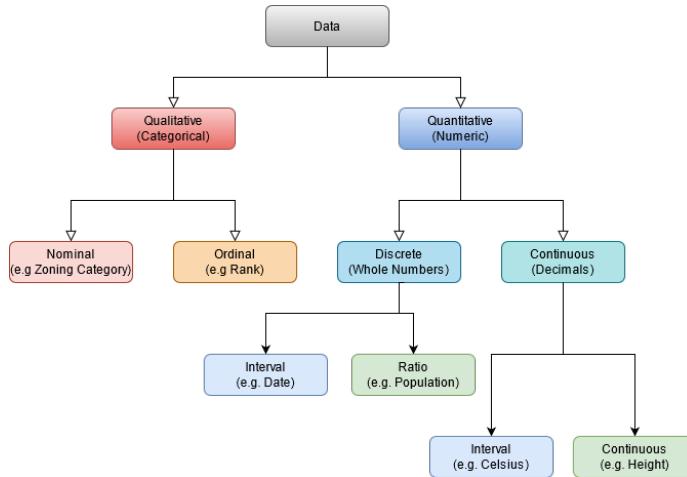


Figure 3.6: Summary of the types of data. Skeeter, CC-BY-SA-4.0.

cannot be used for spatial analysis. Qualitative data can be measured on either a **Nominal** or **Ordinal** scale.

(Nominal scale) These are data that just consist of names or categories with no ranking or direction are nominal. One category is not more or less, better or worse than another, they are just different. A good example would be flower types. Other examples would be zoning categories, colors, flavors of ice cream, place names, etc. With nominal data, you can check for equality between entities and you can count occurrence. These are the only operations we can do. You can't calculate



Figure 3.7: Each flower is different, but no flower is 'more' or 'less' a flower than any of the others. Skeeter, CC-BY-SA-4.0.

(Ordinal scale) These data are categories that also have some ranking or directionality. A good example would be relative sizes (see Figure 3.9). Some other good examples of ordinal data include spice levels (mild, medium, hot), residential zoning density (low, medium, high), and survey responses.

The only arithmetic operations we can do with nominal data are checking for equality (True/False), counting occurrences (frequencies), and calculating the mode (most frequent occurrence). With ordinal data, we can do these operations as well, plus a few more. We can check the order/rank (greater than, less than) and in some circumstances we can calculate the median (see Figure 3.10).

(Graded Membership) When trying to group real world phenomena into categories, there are often “exceptions” that blur the lines a bit. Take this example: you are trying to develop a land cover classification scheme for Garibaldi Provincial park in British Columbia. Some of the land surface is unquestionably alpine tundra and some is certainly forest area. However, the transition between forest and alpine meadow is not an abrupt line. How/where do you draw the line? Examples like this are known as fuzzy variables, and we often use a *Graded Membership* scale to assign them to categories. With the landscape classification, a simple approach would be a “winner take all” approach. If a plot is 5% bare rock, 40% forest, and 45% alpine meadow, the area will be classified as alpine meadow. From that point forward, in the GIS, that area will be treated

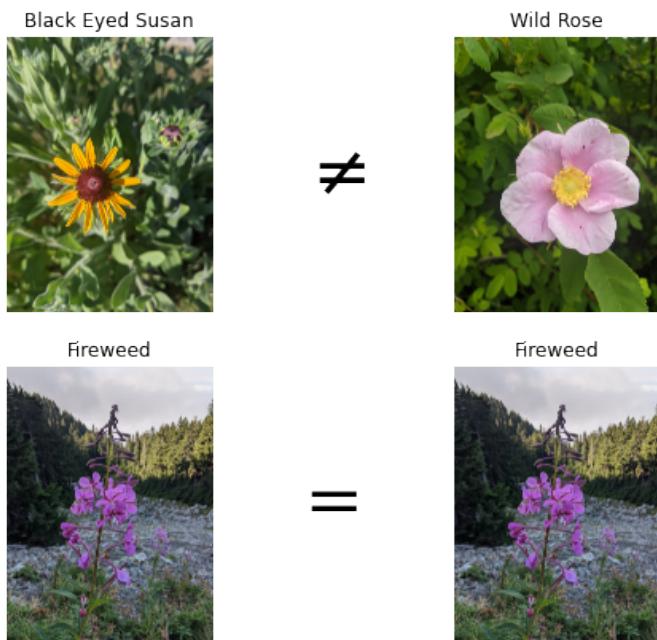


Figure 3.8: Checking equality with flower species. Skeeter, CC-BY-SA-4.0.



Figure 3.9: We can see Yarrow is taller than her sister Shamsa, so we can rank these dogs by height. However, we haven't measured their heights, so we don't know how much taller Yarrow is than Shamsa. Skeeter, CC-BY-SA-4.0.

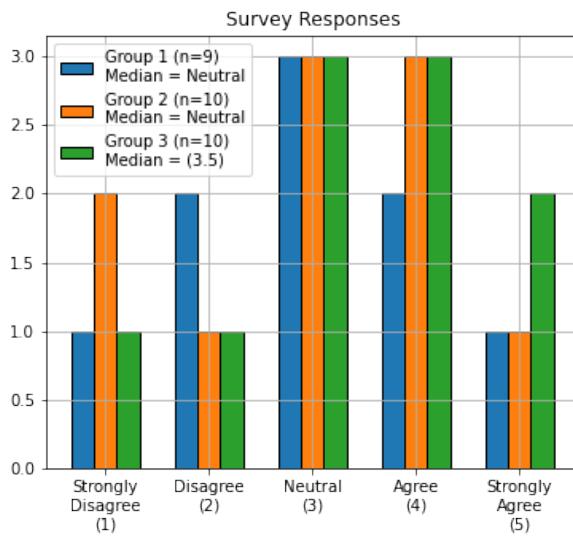


Figure 3.10: In some circumstances, we can directly calculate the median (middle value) of an ordinal set. With odd numbered sets (e.g. Group 1), the median, is simply the middle value of the set, when sorted lowest to highest. We can always take the median when we have an odd number. With even numbered sets, its a bit more complicated. The median, is the average of the middle two values. For Group 2, the middle values (5th and 6th) are both 'Neutral', so we don't have an issue. But for Group 3, the 5th value is 'Neutral' and the 6th value is 'Agree'. We can't directly average these two ordinal values. One solution is to arbitrarily assign a numeric score to the ordinal categories (e.g. 1-5). This would then allow you to show the median is between 'Neutral' and 'Agree'. Skeeter, CC-BY-SA-4.0.

as alpine meadow, any information about the variability within the area will be lost. In practice, many of the qualitative data we work with in GIS, especially those describing natural phenomena, are actually graded membership variables.



Figure 3.11: In this example, we see an alpine landscape in Garibaldi Provincial Park, BC. We can see patches of forest and patches of meadow. But where, exactly, would we draw the boundary between these two landscape classes. Skeeter, CC-BY-SA-4.0.

### 3.2.2 Quantitative Data

Quantitative data are numeric; they describe the quantities associated with an phenomena. The numerical values that are separated by a unit that has some inherent meaning (as opposed to the arbitrary numeric codes like in the ordinal data example). This allows us to conduct a wider range of arithmetic operations on quantitative data. In addition to the operations we perform on Qualitative data; with numeric data we can always calculate measures of central tendency (mean/median) and we can add/subtract values to calculate differences.

Numeric data can be either **discrete** or **continuous**. Discrete variables (e.g. population) are obtained by counting and values within a range cannot be infinitely subdivided. You can have a population of 1, 37, or 179 but you cannot have a population of 2.3. Continuous variables (e.g. temperature) can take an infinite number of values a given range, but they cannot be counted. You can have temperatures of 10, 10.5, or 10.1167 °C, but a temperature of 10°C does not mean you have 10 individual degrees of temperature. Quantitative data (both discrete and continuous) can be measured on either an **Interval** or **Ratio** scale. These types of quantitative data are closely related, but have one important distinction.

Ratio data have fixed, meaningful, absolute zero points. The absolute zero point means ratio data cannot take negative values. It also means that we can multiply/divide two values to calculate a meaningful ratio between them (hence the name). A good example of ratio data are population total (see Figure 3.12). Population counts start at zero and go up from there. A population of zero means there are no residents, and its impossible to have a negative population. Other examples of ratio data include: temperature (*in degrees Kelvin*), precipitation, tree height, income, rental cost, and units of time (years,

seconds, etc.)

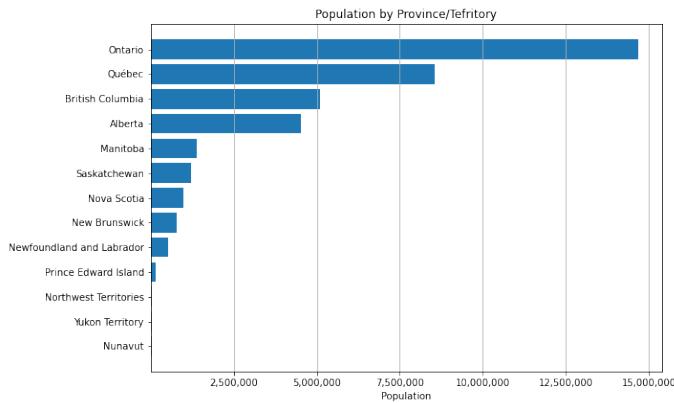
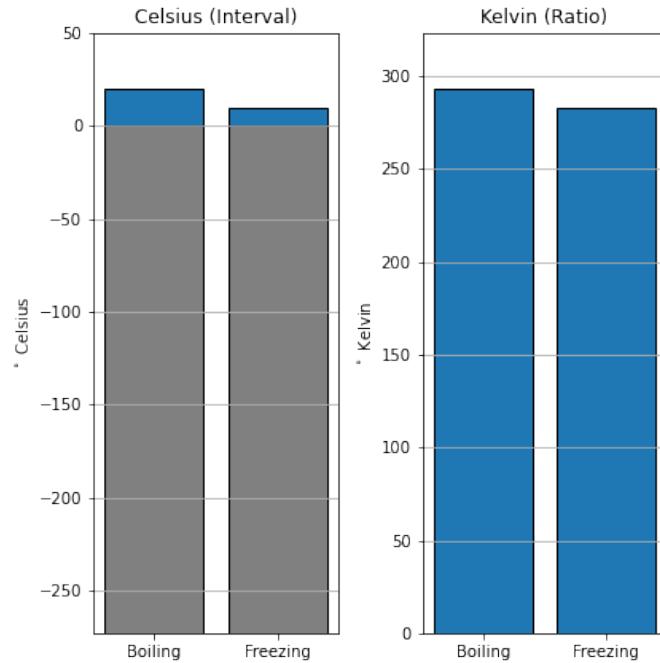


Figure 3.12: Because of the fixed, meaningful zero point, we can calculate ratios between populations: e.g. Manitoba's population is 1/10th that of Ontario, British Columbia has 129 times as many people as Nunavut. Skeeter, CC-BY-SA-4.0.

Interval data on the other hand, have an arbitrarily set zero point. This means they can take negative values. Because the zero point is arbitrary, we cannot multiply/divide two values or calculate meaningful ratios between two values. A good example of interval data is temperature measured in Celsius, and comparing it to Kelvin highlights the difference between the two data types (see Figure 3.2.2). The conversion between Kelvin (ratio) and Celsius (interval) is very simple:  $^{\circ}\text{C} = ^{\circ}\text{K}-273.15$ . Zero Kelvin is “Absolute Zero” - ie. the lack of temperature, while zero Celsius is the freezing point of water (273.15 degrees above absolute zero). Other examples of interval data include: the pH scale, IQ test scores, elevation (relative to a datum) dates (April 12th, 2011), and times (11:00 A.M.).



\begin{figure}

\caption{The ratio between two temperatures in Celsius is not meaningful, 20°C is not ‘twice’ as warm as 10°C. Kelvin’s zero point is fixed to absolute zero, the ‘absence’ of temperature. So we can calculate the ratio, 293.15°K is 1.035 times warmer than 283.15°K. Skeeter, CC-BY-SA-4.0.} \end{figure}

### 3.2.3 Derived Ratio: Normalizing Data

Sometimes we want to account for the influence of one variable when analyzing another. To do this, we can divide one value by another to get the ratio of the two, also known as a **derived ratio**. This process is sometimes referred to as **Normalizing** or **Standardizing** our data. The basic formula is:  $C = \frac{A}{B}$ , where A is our variable of interest, B is our confounding variable, and C is our new derived ratio. There are many circumstances where we might need to do this. One common example is population density: Canada and Poland both have populations of  $\sim 38$  million people but Canada had 32x the land area of Poland. Any comparison of the these two nations that fails to account for the size disparity would be seriously flawed. Another key example are affordability indexes. The example below shows how normalization can be applied to a households expenditures on food. Income and household expenditures on food are strongly related (wealthy regions tend to purchase more expensive food). An analysis of the cost of food that doesn’t account for this relationship would not adequately account for the *affordability* of food in a given region. Dividing household food expenditures by household income, we get the proportion of income spent on food. This is a much more accurate representation of the afford-

ability of food and highlights that the poorest communities are most severely impacted by increasing food costs.

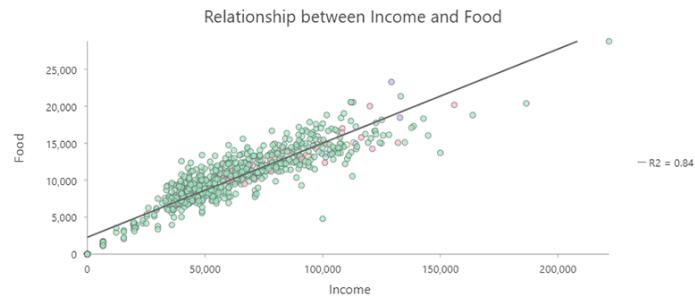


Figure 3.13: Income vs. household expenditures on Food by Census Subdivisions in BC. Skeeter, CC-BY-SA-4.0.

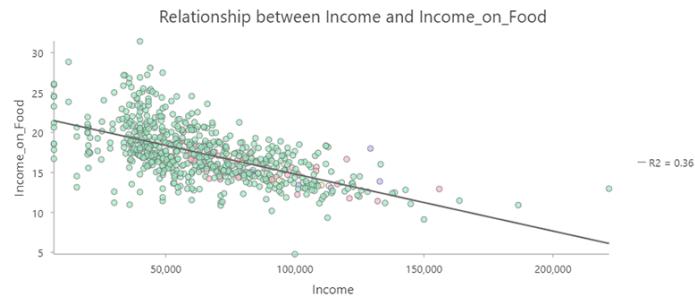


Figure 3.14: Income vs. Fraction of Income Spent on Food by Census Subdivisions in BC. Skeeter, CC-BY-SA-4.0.

### 3.2.4 Summary of Data Types

Operation	Nominal	Ordinal	Interval	Ratio
Equality	x	x	x	x
Counts/Mode	x	x	x	x
Rank/Order		x	x	x
Median		x	x	x
Add/Subtract			x	x
Mean			x	x
Multiply/Divide				x

### 3.3 Spatial Is Special

You might encounter the phrase “Spatial is special” in your time studying GIS. Spatial data is the foundation of Geographic Information Science, it is what distinguishes GIS from the broader field of data science. This was succinctly summarized by Waldo Tobler in The First Law of Geography: - *“Everything is related to everything else, but near things are more related than distant things.”*

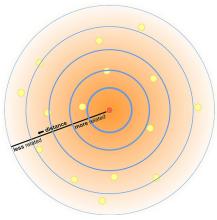


Figure 3.15: Visualization of Tobler’s First Law. Skeeter, CC-BY-SA-4.0.

This might seem obvious: people interact more if they live in the same city, orca pods in different areas develop different dialects, hemlocks on Vancouver Island are more related to their neighbors than to hemlock in the New Brunswick. Generally, near things are more related to one another, but it *does not guarantee similarity*. Downtown Vancouver averages 40 cm of snow/year, but the ski resort on Grouse Mountain 15 km north gets over 9 m. These locations are impacted by the same storm systems, but the 1200 m elevation difference causes vastly different quantities and different types of precipitation.

The measure of similarity between objects across space called **spatial autocorrelation**. Spatial autocorrelation allows us to make some key assumptions when representing spatial data. We don’t have to measure a phenomena everywhere in order to represent it adequately. We only need to measure it at specific locations or over regular intervals. If point A is in dense forest, it is likely point B 10 m away is also in a dense forest. We don’t have to get the location of every tree in the forest. Instead, we can look at the average presence of trees over a larger area.

### 3.4 Spatial Data Models

As discussed in the previous chapter, spatial data is three-dimensional, though we usually project it into two-dimensions for simplicity. Because of the unique transformations that must be applied to spatial data, it must be treated and represented differently than the non-spatial data that describe *what* is happening and *when*. We can’t simply put all of our data into a spreadsheet and start analyzing it. We have to use **Spatial Data Models** to organize our data and link our spatial and non-spatial data. Spatial data models store geographic data in a systematic way so that we can effectively display, query, edit, and analyze

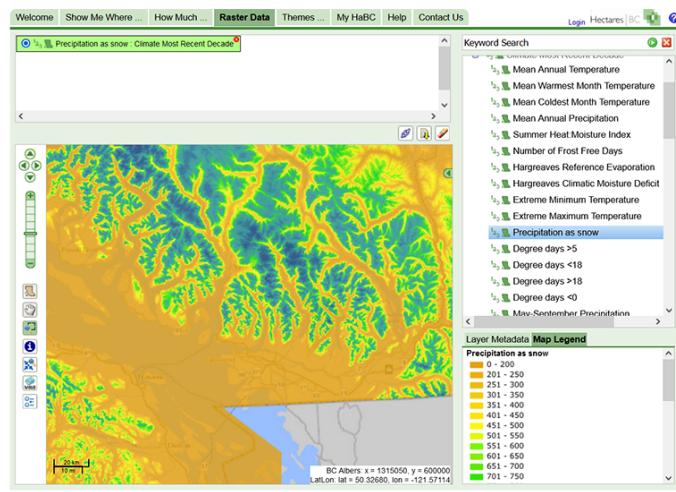


Figure 3.16: Screengrab from Hectares BC, can easily make into better map.  
Skeeter, CC-BY-SA-4.0.

our data within a GIS.

There are two main types of spatial data models: the **Raster** and **Vector** models. The raster data model represents spatial data as grid of cells, and each cell has one non-spatial attribute associated with it. The vector data model represents spatial data as either points, lines, or polygons that are each linked to one or more non-spatial attributes. These two models represent the world in fundamentally different ways. One is not inherently better than the other, but they are better suited for different circumstances. The choice of which model to use is often dictated by three main factors:

- 1) The type of phenomena we are trying to represent.
- 2) The scale at which we plan to analyze our data.
- 3) How we plan to use the data.

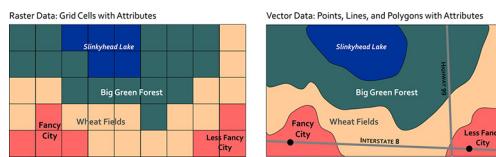


Figure 3.17: Representing space in the raster model vs. the vector model.  
Skeeter, CC-BY-SA-4.0.

### 3.4.1 Raster Data Model

The raster data model represents a phenomena across space as a gridded set of cell (or pixels). The cell size determines the **Resolution** of the raster image, that is the smallest feature we can resolve with the raster. A 10 m resolution raster has cells that are 10 x 10 m (100 m<sup>2</sup>), a 2 m resolution has cells that are 2 x 2 m (4 m<sup>2</sup>). Along with the cell size, the number of rows and columns dictates the extent (or bounds) of a raster image. A raster with a 1 m cell size, 5 rows, and 5 columns, will cover an area of 5 m x 5 m (25 m<sup>2</sup>). Because of the full coverage within their bounds, raster data models are very well suited for representing *continuous phenomena* where cell values correspond to measured (or estimated) value at specific location. In GIS, rasters are commonly encountered as: satellite and drone imagery, elevation models, climate data, model outputs, and scanned maps.

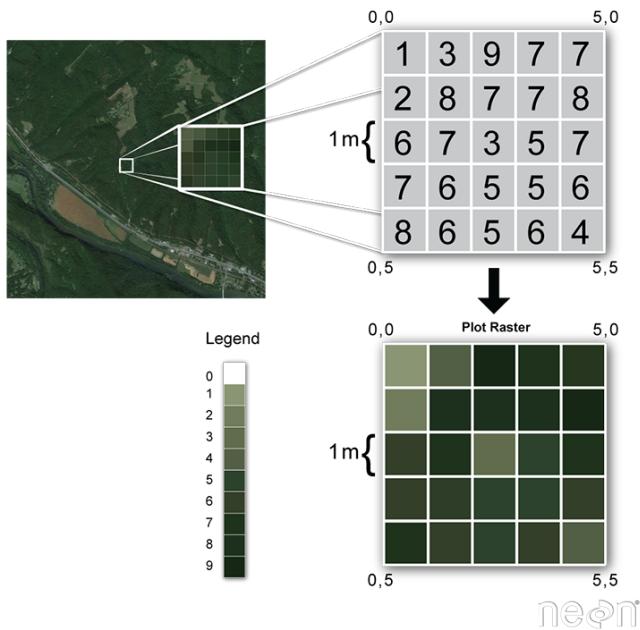


Figure 3.18: I think this one is pretty good, I use it in my lecture notes. Skeeter, CC-BY-SA-4.0

The value of a pixel can be quantitative (e.g. elevation) or qualitative (e.g. land use). Each pixel/cell can only have a single value associated with it. Multiple bands can be combined to store or more information, as is done with a RGB color photograph. Algebraic expressions can also be performed quickly and efficiently with raster layers as inputs. This is known as raster overlay, and is one of the key advantages to raster data. If layer A = Average July Temperature and layer B = Average January Temperature, then A - B will give us the Average

Temperature Range across the rasters domain.

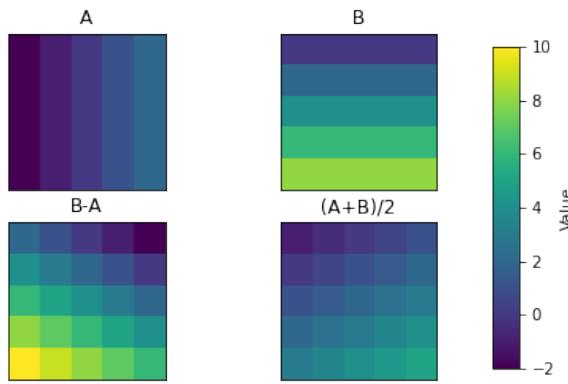


Figure 3.19: Raster math illustration. Skeeter, CC-BY-SA-4.0.

Rasters data relies on Spatial Autocorrelation and The First Law of Geography, the model assumes that *all areas* within a given cell are equally represented by the cell value. Depending on the resolution of the raster and the scale of the task at hand, this may or may not be an effective assumption. If you are trying to represent the coastline of Nova Scotia, 100 m or even 1 km resolution cells will likely suffice (see Figure 3.20). However, 10 km cells severely degrade the quality of the representation and at a 100 km cell size, the province is indistinguishable.

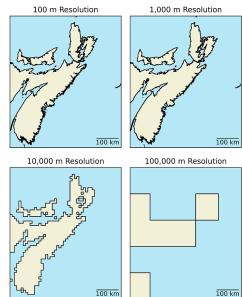


Figure 3.20: Raster Resolution. Skeeter, CC-BY-SA-4.0.

The above example is related to something known as them *mixed pixel* problem. Each cell in a raster can only have **one** value. So how do you handle when the area a cell covers contains **multiple** examples. Possible approaches are:

- **Majority/Mode**, the cell value is determined by the value/class covering the largest area within each cell. This can be useful for **discrete**

phenomena, but generally won't be helpful for continuous phenomena.

- **Touches All**, can be useful for discrete phenomena if you need to prioritize specific class(es) you can designate it them to be assigned to any pixel the touch (eg. with flood/fire risk or most other hazards, you want to take an inclusive approach when defining risk zones. Better safe than sorry)
- **Nearest Neighbor/Center Point**, the cell value is determined by the value/class only at the center point of the cell. This method is quick to calculate but can under/overestimate repeating phenomena with frequencies lining up with the raster resolution (eg. City Blocks/Roads, rows in agricultural fields)
- **Average**, when working with continuous phenomena (eg. rainfall, temperature, elevation) it might be best to use the average value across the cell instead. If multiple observations are available calculate the spatially weighted average within each cell. If we are working with discrete phenomena, this method is generally less useful.

When the data resolution is very high, relative to the scale of the map/analysis, the specific choice of method will produce negligible differences. If you're working with a 25m resolution land cover classification and doing a continental scale analysis, the improper attribution of boundary pixels won't have a huge impact on the results. If the data resolution is low relative to the scale of your analysis, the choice of method could have a significant impact on your results.

Raster data can come in many different formats. **GeoTIFF** which has the extension .tif is one of the most common/functional is the . This format is based on the Tag Image File Format (TIFF), a common file type used by graphic artists and photographers. A TIFF file stores metadata (data about the data) as tags. For instance, your camera might store a tag that describes the make and model of the camera and another for the date the photo was taken when it saves a picture. A GeoTIFF is a standard .tif image format plus additional tags spatial tags denoting spatial information including:

- Extent (minimum x,y and maximum x,y)
- Resolution (cell size)
- Projection, Coordinate system, and datum

Other file types you will likely encounter when working with raster data include:

- 1) IMG - A proprietary image format commonly used by ESRI products
- 2) JPEG2000 - A geospatial version of the common .jpg image type
- 3) ASCII - An older human readable format (simple text file) with slower performance than the types listed above.

### 3.4.2 Vector Data

The vector data model is much more well suited to represent discrete phenomena than the raster data model. A vector feature is a representation of a discrete object as a set of x,y coordinate pairs (points) linked to set of descriptive attribute about that object. A vector feature's coordinates can consist of just one x,y pair to form a single point feature, or multiple points which can be connected to form lines or polygons (see Figure 3.21). The non-spatial attribute data is typically stored in a **Tabular** format separate from the spatial data, and it is linked using an index. One of the key advantages of the vector model is the ability to store and retrieve many attributes them quickly. In GIS, vector data are commonly encountered as: political boundaries, census data, pathways (road, trails, etc.), point location (stop sign, fire hydrant), etc.

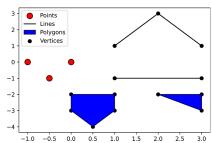


Figure 3.21: Vector objects (points, lines, or polygons) are stored along with any number of attribute. Point, line, and polygon data are typically stored in separate files. Skeeter, CC-BY-SA-4.0.

**Points** are “zero-dimensional”, they have no length, or width. A point feature is just an individual , coordinate pair representing a precise location, that has some linked attribute information. Points are great for representing a variety of objects, depending on the scale. Fire hydrants, light poles, and trees are suitable to be represented as points in almost any application. If you are making a map of mines in British Columbia, or cities across Canada, it’s probably acceptable to just display them as points.

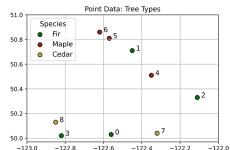


Figure 3.22: An example of point data showing locations of trees. The points are labeled with their index (unique ID number) which corresponds to the attribute table below which stores more information about each tree. Skeeter, CC-BY-SA-4.0

index	Longitude (X)	Latitude (Y)	Name	Age	Height
0	0.44	0.03	Fir	54	119
1	0.55	0.71	Fir	29	56
2	0.89	0.33	Fir	82	197
3	0.18	0.02	Fir	46	98
4	0.65	0.51	Maple	87	212
5	0.43	0.81	Maple	73	172
6	0.38	0.86	Maple	94	233
7	0.68	0.04	Cedar	34	68
8	0.15	0.13	Cedar	36	73

**Lines** are one-dimensional, they have length, but no width and thus no area. A line consists of two or more points. Every line must have a start point and end point, they may also have any number of middle points, called vertices. A vertex is just any point where two or more lines meet. Lines are also great for representing a variety of objects, depending on the scale. Hiking trails, flight paths, coastlines, and power lines are suitable to be represented as lines in almost most applications. When making smaller scale maps, its often sufficient to represent rivers as lines, though at large scales we might elect to use a polygon.



Figure 3.23: Roads are typically reprented as line data. Though they obviously have an area, unless we are making a very large scale map, we don't need (or have the room) to show that on a map. This BC road altas makes use of line data, representing roads a lines and using different colors to denote the type of road. Skeeter, CC-BY-SA-4.0.

**Polygons** are two-dimensional, they have both a length and width and therefore we can also calculate their area. All polygons consist of a set of at three or more points (vertices) connected by line segments called “edges” that connect to form an enclosed shape. All polygons form an enclosed shape, but some can also have

“holes” (think doughnuts!), these holes are sometimes called interior rings. Each interior ring is a separate set vertices and edges that is wholly contained within the polygon and no two interior rings can overlap. Polygons are useful for representing many different objects depending: political boundaries boundaries, Köppen climate zones, lakes, continents, etc. At large scales they can represent things like buildings which we might choose to represent as points at smaller scales.

Sometimes, a discrete object has multiple parts, that are spatially separated. In these circumstances, the vector model allows for multi-polygon, multi-line, or multi-point objects. A good example of when a multi-polygon would be useful is the StatsCanada provincial boundary file (see Figure 3.24). Roads sometimes need to be stored as multi-lines as well, for example Highway 1 crosses the Georgia Straight from Vancouver to Nanaimo. If we want the to represent the entire Highway as one object, we need to use a multi-line.



Figure 3.24: This is the official Stats Canada provincial boundary layer. All the other coastal provinces and territories have islands. We don’t need to represent every island as a separate object, so we can ‘bundle’ together the polygons as multipolygons. The landlocked provinces do not have any coastlines and are represented as simple polygons reather than multipolygons. The attribute table below corresponds to the map and lists the geometry type (polygon/multipolygon). Skeeter, CC-BY-SA-4.0

PRNAME	Province ID	Population	Area	Geometry Type
Newfoundland and Labrador	10	525572	373872	MultiPolygon
Prince Edward Island	11	157329	5660	MultiPolygon
Nova Scotia	12	971451	53338	MultiPolygon
New Brunswick	13	779940	71450	MultiPolygon
Quebec	24	8536855	1365128	MultiPolygon
Ontario	35	14666590	917741	MultiPolygon
Manitoba	46	1389952	553556	MultiPolygon
Saskatchewan	47	1206019	591670	Polygon
Alberta	48	4511223	642317	Polygon
British Columbia	59	5111756	925186	MultiPolygon
Yukon	60	41774	474391	MultiPolygon
Northwest Territories	61	45217	1183085	MultiPolygon
Nunavut	62	39419	1936113	MultiPolygon

Vector data also has a **Resolution** although it has a somewhat different definition in the context of the vector model. Vector resolution is determined by the smallest resolvable feature. Another way to describe vector resolution, would be the distance between vertices. The greater the distance between vertices, the fewer vertices there are per polygon and, the lower the resolution. If a vector object (line or polygon) has many vertices, we will have a higher resolution representation of the feature.

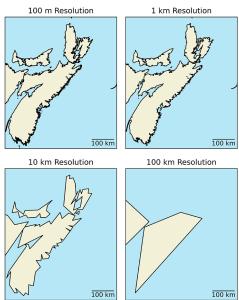


Figure 3.25: Vector image of Nova Scotia at different resolutions. Here the original polygon (top left) has been downsampled to lower resolutions, by setting the minimum allowable distance between vertices. As the distance between vertices increases, the resolution decreases and the coastline becomes less distinguishable. Skeeter, CC-BY-SA-4.0.

Like raster data, vector data can also come in many different formats. The **shapefile** format which has the extension .shp is one of the most common file types you will encounter. A .shp file stores the geographic coordinates of each vertex in the vector, as well as metadata including:

- The spatial extent of the shapefile (i.e. geographic area that the shapefile covers). The spatial extent for a shapefile represents the combined extent for all spatial objects in the shapefile.
- Object type - whether the shapefile includes points, lines, or polygons.
- Coordinate reference system (CRS)
- Attributes - for example, a line shapefile that contains the locations of streams, might contain the name of each stream.

Because the structure of points, lines, and polygons are different, each individual shapefile can only contain one vector type (all points, all lines or all polygons). You will not find a mixture of point, line and polygon objects in a single shapefile.

**GeoJSON** is a simple, lightweight format for storing a variety of geographic data structures. It is most commonly encountered in web mapping and other open source applications. GeoJSON supports the following geometries: Point, Line, Polygon, MultiPoint, MultiLine, and MultiPolygon objects. Unlike with shapefiles, one GeoJSON file can contain any mix of geometries. An objects with and its attributes are a Feature object. A set of Features is a FeatureCollection. GeoJSON has the added benefit of allowing you to encode stylistic choices within the file. If you'd like to explore this format a bit more, copy the code below and paste it in the online GeoJSON editor geojson.io. You can make changes and see them reflected on your the map.

```
{
  "type": "FeatureCollection",
  "features": [
    {
      "type": "Feature",
      "properties": {
        "marker-color": "#blue",
        "marker-size": "medium",
        "marker-symbol": "circle",
        "Name": "Vancouver"
      },
      "geometry": {
        "type": "Point",
        "coordinates": [
          -123.04687499999999,
          49.23912083246698
        ]
      }
    },
    {
      "type": "Feature",
      "properties": {
        "marker-color": "red",
        "marker-size": "medium",
      }
    }
  ]
}
```

```

    "marker-symbol": "square",
    "Name": "Victoria"
},
"geometry": {
    "type": "Point",
    "coordinates": [
        -123.40942382812501,
        48.516604348867475
    ]
}
}
]
```

**Simple text files** are human readable file formats (.txt, .csv) that are suitable for storing point and attribute data. You will often encounter .txt or .csv files when working with weather data for instance (see Table). Coordinates (typically latitude and longitude) are stored in a text files along with the other attributes. We can bring this type of file into a GIS, but we need to convert the data to point features before we can display it.

#### *Canadian Weather Station File*

Name	Province	Climate ID	Latitude (Decimal Degrees)	Longitude (Decimal Degrees)
ACTIVE PASS	BRITISH COLUMBIA	1010066	48.87	-123.28
ALBERT HEAD	BRITISH COLUMBIA	1010235	48.40	-123.48
BAMBERTON	BRITISH COLUMBIA	1010595	48.58	-123.52
OCEAN CEMENT	BRITISH COLUMBIA			
BEAR CREEK	BRITISH COLUMBIA	1010720	48.50	-124.00
BEAVER LAKE	BRITISH COLUMBIA	1010774	48.50	-123.35
BECHER BAY	BRITISH COLUMBIA	1010780	48.33	-123.63
BRENTWOOD BAY 2	BRITISH COLUMBIA	1010960	48.60	-123.47
BRENTWOOD CLARKE ROAD	BRITISH COLUMBIA	1010961	48.57	-123.45

Name	Province	Climate ID	Latitude (Decimal Degrees)	Longitude (Decimal Degrees)
BRENTWOOD	BRITISH COLUMBIA	1010965	48.57	-123.43
W SAANICH RD	COLUMBIA			
CENTRAL SAANICH VEYANESS	BRITISH COLUMBIA	1011467	48.58	-123.42

## 3.5 Choice of Spatial Data Model

There is no “best” spatial data model. Rasters are more well suited for some applications and vector data are better suited for others. The section summarizes some of the key considerations that influence which model is suited for which situations.

### 3.5.1 Comparing Data Models

Vector	Raster
Usually discrete objects Points, Lines, and/or Polygons	Usually continuous fields Grid of cells (pixels) with continuous coverage
Each object can have many attributes	Each cell has one value per band (layer)
Objects may overlap, have gaps, or be continuous	One raster image can have many bands

### 3.5.2 Raster Data Model

Advantages	Disadvantages
Well suited for continuous variables: in space and time Simple data structure makes overlay is easy and efficient	Large file size: exponentially proportional to resolution and linearly proportional to number of bands. Loss of information during rasterization (mixed pixel problem, see case study). Reductions in cell size may lead to inability to recognize spatial features.

### 3.5.3 Vector Data Model

Advantages	Disadvantages
Compact data structure: smaller file sizes	Complex data structures compared to rasters
A good representation of discrete objects	Topology (connectivity) - can be a huge head ache when creating a layer
Easy to query and select by attributes	Some tasks (overlay of layers) can be computationally expensive
Graphic output is usually more aesthetically pleasing	No variability within polygons possible
Topology (connectivity) - Proximity & Network Analysis	Less suited for continuous variables (requires significant generalization) or temporal change

### 3.5.4 Which Data Model is Best?

No single data model is suitable for all types of data or analysis.

- Most GIS systems employ both raster and vector data structures so that the user can choose the model best suited to the representation of their data
- It is possible to convert back and forth between models
- However, this results in a loss of information and may introduce additional error each time a conversion is made

## 3.6 Case Study

### 3.6.1 Resolution: Vector vs. Raster

This video gives an applied example of how resolution differes between raster and vector data. *Note* The video is just a placeholder.

#### Your turn!

I'll do some exercise building on the case study.

#### Call out

This is a call out. Put some important concept or fact in here.

## 3.7 Summary

This chapter has introduced you to how we represent data in a GIS.

1. Types of Phenomena
2. Types of Data

3. Raster Data Models
4. Vector Data Models
5. Data Resolution

### **Reflection Questions**

1. Explain the difference between continuous fields and discrete objects.
2. Define Quantitative data and Qualitative data.
3. What is the role of resolution in raster data?
4. How does the vector data model differ from the raster data model?

### **Practice Questions**

1. Given ipsum, solve for lorem.
2. Draw ipsum lorem.

Ensure all inline citations are properly referenced here.

## **Chapter 4**

# **Collecting and Editing Data**

Written by Ira Sutherland and Paul Pickell

The ability of a geomatician to answer research questions or produce a map or other visuals, rests, in part, on first finding the right data to do so. Geomaticians often spend much of their time finding, collecting, and editing data, yet the critical activity of finding data are often left as something that geomaticians are assumed to pick up along the way. This chapter addresses that gap by first introducing a range of possible data sources along with some theory, tips, and strategies to access them. We also address some common instances when data do not yet exist, and so we must create them. This chapter may be particularly useful for students and researchers starting out on their spatial research projects, and for anyone interested in the rapidly changing data universe.

### **Learning Objectives**

1. Become familiar with a wide range of spatial datasets and strategies to access them
2. Identify several sources of historical spatial information, including historical maps and aerial photos, and the steps required to analyze them as spatial information
3. Recognize good practices and strategies for writing and reading metadata
4. Understand the components of Global Navigation Satellite Systems (GNSS) and how location on Earth is triangulated from these systems

### **Key Terms**

Aerial Photography, Area of Interest, Census, Data Repository, Data Request, Georeferencing, Global Navigation Satellite Systems (GNSS), Natural Resource

Administrative Data, Historical Collections, Open Data, Orthophoto, Spatial Panel Data

## 4.1 Open Data

Data are becoming increasingly easy to access thanks to the open data movement. The concept of **open data** suggests that governmental data should be available to anyone to use and, if desired, redistribute in any form without any copyright restriction (Kassen 2013) or with minimal restrictions such as providing recognition [Kassen, 2013].

Until recently, most government data were simply unavailable or could only be accessed by data request or by paying the government data provider. Countries around the world are moving to an open data model. For example, Britain is opening up its national geographic database (housed as the ‘Ordnance Survey’). United States (US) has moved its data housed within the US Geological Survey into the public domain [USGS]. Canada has signed a Directive on Open Government, which promotes the proactive and ongoing release of government information. The province of British Columbia (BC) has just released a large collection of provincial LiDAR data under an open government license and many provinces and municipalities release data under similar licenses. Canada is also signatory of the Treaty of Open Skies, which is an international effort that encourages the sharing of aerial imagery to promote openness and transparency of each signatory nation’s military forces and activities. Despite the tremendous momentum towards open data, many datasets are not yet fully open. The tips and strategies below will help locate both open and not-so-open datasets.

## 4.2 Finding Data

Here we introduce a network model to set a framework for finding data. Imagine that nearly all the data and information in the world is connected in some way through networks of information, composed of individuals, libraries, and institutions. The internet is an important component in this network, one we all use every day to answer questions. For example, me might ask Google: “what is the best lake in Canada to plan a summer holiday?” A common answer returned is ‘Lake Louise, Alberta,’ which is a stunning lake surrounded by tall Rocky Mountains, as well as hordes of tourists! If we asked this question to our friends – and maybe one happens to be an expert fisherman or fisherwoman - we may receive different answers including secret lakes that have not yet been discovered by tourists, or the best lake for fishing. Our friends can also consider our specific interests, suggest helpful resources (such as a lesser known forum on local fishing), and offer additional information about our query such as the best places on that lake to camp, where to fish on the lake, and what type of fishing gear to use. The point in this example is that there are different networks of information available to us, including formal networks of information organized on the

internet and accessed by search engines as well as informal networks of individuals and experts who offer an additional strategy to connect us with the right information. Data are becoming increasingly easy to discover through the use of **data repositories** (Figure 4.1). Below we discuss the growing number (and centralization) of spatial data repositories, which can give access to academic, government non-governmental, international, and crowdsourced datasets. Here we introduce each type of repository and offer some hints at what environmental data can be discovered in each.

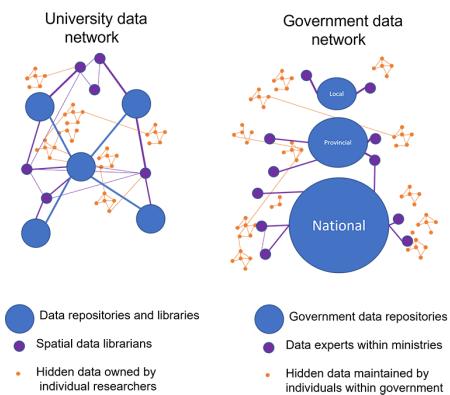


Figure 4.1: Envisioning university and government data networks. Data within each network is concentrated within data repositories, yet considerable data remains 'hidden' among individual researchers and silos of the Ministry, but can potentially be accessed by finding the right connections. Sutherland, CC BY 4.0.

## 4.3 Data in Academia

Data librarians are particularly well connected and trained to help you navigate these repositories and contacting them can be a good starting point in your search. Nonetheless, considerable data are not yet published. Some of this unpublished data has been analyzed in previous research and its existence could be discovered through a review of the academic literature. Other unpublished data remains essentially 'hidden,' only known about by individuals or small clusters of individual researchers who created those data. In such a case, your only possibility to find such data are through a combination of 'asking around' and reaching out to experts in the field. Once you know it exists, unpublished data could potentially be accessed through connecting with those researchers themselves, and requesting the data or inquiring about the possibility for a collaboration.

## 4.4 Government Data

Government data are also increasingly published in data repositories, specific to the level of government (Figure 4.1). There are multiple levels to government, including municipalities (the smallest), provinces (or states), and nations (the largest), each of which often has its own data repository. Centralized repositories are becoming increasingly common and connect open data from all levels of Government. The Federated Research data repository is an aggregation of Canadian open data repositories, including municipal, provincial, and academic repositories. It includes a map-based search for datasets with location information tied to their metadata. In the US, geospatial data from federal, municipal, and state government repositories are being consolidated under Data.gov.

Because not all repositories are yet connected by a centralized repository, one must search in the correct repository. To do this, consider which government has jurisdiction over the specific subject area and geography of interest. For example, if you are interested in land use zoning and engineering features within a given city, this data are likely best provided by that individual city, either by finding it within a data repository or emailing the municipality with a data request (discussed below). In Canada, the provinces have jurisdiction over most natural resources and thus provincial government data repositories tend to provide data on natural resources, such as, water features, forests, wildlife, minerals, and topography. In British Columbia, for example, DataBC houses over a thousand datasets on natural resources, including forest cover mapping, natural disturbances, hunting statistics, administrative boundaries, and much more. Canada's open data portal provides data on fish as well as environmental conditions (e.g., water quality, air quality, historical weather, etc.), which is under federal jurisdiction. Hydrological flow and water quality monitoring is readily accessible across Canada through the Hydat database, which can be easily accessed through the R package TidyHydat (Albers 2017).

### Your Turn!

Try using a web search to find the government open data pages for your city, province/state, and nation. What kinds of data can you find?

## 4.5 Census Data

This section introduces the census at a cursory level before launching into the applied question of how to find census data for your spatial analysis, using the Census of Canada as an example.

A **census** generally refers to a complete count by government of a specific region's population by age, gender, language, income, housing and other demographic characteristics. Census data inform public policy, such as allocation of public funds, transportation network planning, and electoral area delineation.

Census data also provide researchers with an opportunity to gain insight into the social and, to a lesser extent, environmental fabric of a country and are increasingly used in environmental and social-ecological research that aims to address social elements of environmental challenges [Tomscha et al., 2016] [Biggs et al., 2021]. Censuses are typically conducted once every five years (e.g., Canada) or every 10 years (e.g., United States).

In addition to demographics, many nations survey information related to economics or specific industries, such as agriculture. For example, Canada's Census of Agriculture captures information on fertilizers, irrigation, livestock, farm types, and crop production across Canada. The Longform Census in Canada surveys additional questions but is only sent to a subset of the population, and the data from it are then estimated for the entire population.

A starting point to using census data in spatial analysis is to understand the geographic levels of census data, and then we address where the geography files and data can be downloaded.

## 4.6 Census of Canada Geographic Levels

To protect respondents' confidentiality, the individual data collected during census enumeration is obscured from the public. Thus, census data can only be accessed by researchers in the form of statistics aggregated to varying geographic levels. Knowing these geographic levels is key to accessing census data.

At the top of Figure 4.2 are Canada's provinces and territories, which are then divided into census divisions, which in turn are divided into census subdivisions. Census subdivisions correspond to municipalities, but also include Indian reserves, and 'unorganized areas.' These three areas (municipalities, Indian reserves, and unorganized areas) are also aggregated into census consolidated subdivisions, which offer a more consistent geographic unit for mapping large areas as compared to subdivisions themselves. Census subdivisions are divided into dissemination areas, composed of one or more 'dissemination area blocks' (generally, a city block bounded by roads on all sides).

In addition to these general geographies, which apply throughout Canada, special geographic units are implemented as an additional layer of aggregation for urban centers. A Census Metropolitan Area (CMA) is a grouping of census subdivisions comprising a large urban area and its surroundings. To become a CMA, an area must register an urban core population of at least 100,000 at the previous census. A Census Agglomeration (CA) is a smaller version of a CMA in which the urban core population at the previous census was greater than 10,000 but less than 100,000. CMAs and CAs are useful for making comparisons across cities. CMAs and CAs with a population greater than 50,000 are subdivided into census tracts which have populations ranging from 2,500 to 8,000 and are intended to be relatively homogeneous in their demographic identity (i.e., a local neighbourhood). Using census data for geographic analysis

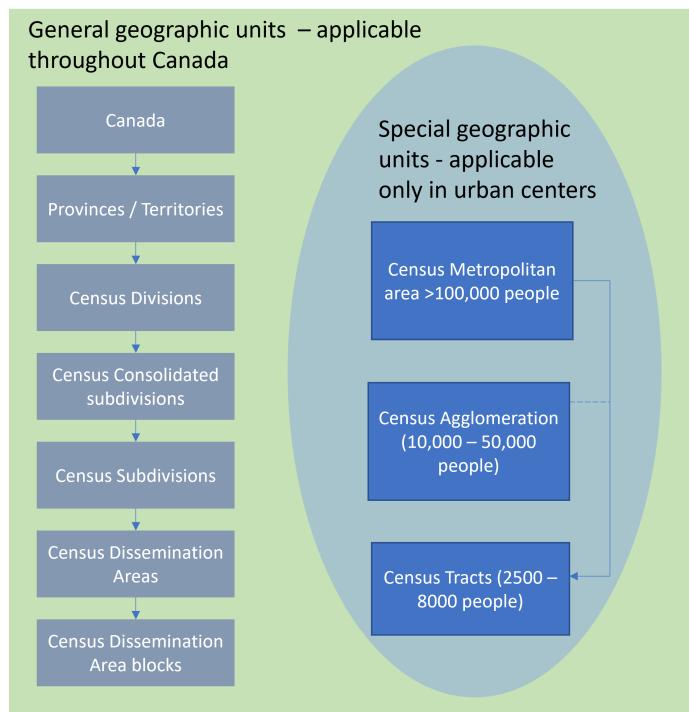


Figure 4.2: The geographic levels of the Census of Canada include general units (applicable everywhere throughout Canada) and also an additional layer for urban areas only. Sutherland, CC BY 4.0.

typically involves first identifying the smallest spatial unit at which the data are available. Recall that to protect the privacy of respondents, some data are only available at higher geographic levels. Another consideration is that if you plan to compile multiple census years, the geographic boundaries have typically changed over time in response to how the landscapes and information needs have changed. This creates substantial (though, not insurmountable) additional work that limits how the data can be used, especially for finer spatial scale analysis. An example of changes in the geography of census divisions is seen for British Columbia in Figure 4.3.

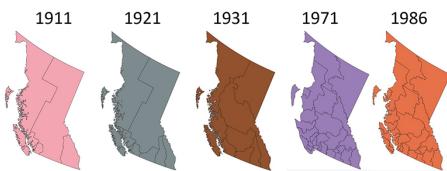


Figure 4.3: An example of how census boundaries have changed, showing changes in Census divisions for British Columbia from 1911 to 1986. Data from Clark [2016]. Sutherland, CC BY 4.0.

## Call Out

Spatial analysis will often want to work with the smallest geographic level available. The smallest geographic unit of the Canada census is the dissemination block. Census tracts are also used frequently in spatial analysis but this geographic unit is only applicable to metropolitan areas.

## 4.7 Accessing Census Data

Statistics Canada maintains the geographic boundaries for the Census for each level in Canada. The Canadian Socio-economic Information Management System Statistics Canada data portal provides access to the Census of Canada as well as the Census of Agriculture, Aboriginal Peoples Survey, and other government statistical datasets. You have the option to search by a vector or an area of interest. Students with access to CHASS Canadian Census Analyzer (students of University of Toronto as well as many other subscribing universities) can use CHASS to access additional statistical data, which they can aggregate to census geographic units of their choosing.

### Your Turn!

Try this: Navigate to Canadian Socio-economic Information Management System Statistics Canada data portal and search a key word such as: “age.” A list of available geographic levels should be present on the left side, allowing you

to check which geographic levels you would like to retrieve the data for. What geographic levels are present for age and which is the smallest geographic level (refer to Figure 4.3)? Now try searching the keyword: crop production. What is the smallest geographic level for crop production now?

## 4.8 Non-Governmental Organization Data

Many elements of the environment, such as biodiversity and large old trees, are not monitored by most governments. These knowledge gaps are sometimes filled by other organizations not associated with the government (i.e., non-governmental organizations) or by citizen science initiatives. For example, Pacific salmon have been a top conservation concern lacking data in western North America. An organization called the Pacific Salmon Foundation has collaborated with the help of First Nations and government to compile salmon information for BC so that the data can be readily viewed and downloaded for further analysis. Organizations such as the International Union for Conservation of Nature often synthesize and offer datasets that support their mandates such as monitoring species at risk and expanding protected areas.

## 4.9 Citizen Science

Citizen science describes activities where members of the general public contribute information and data to help generate new knowledge and information [Lee et al., 2020]. Citizen science has been used to fill in data gaps for widely distributed phenomenon that are otherwise difficult to gather. In addition to Open Street Map, which has created a free open geodatabase of the world, one of the most famous examples is a collective global effort to map the distribution of global bird species, which through the online application called E-bird has generated nearly 1 billion bird observations as of 2021. Likewise, alpine wildlife are difficult for researchers to observe and are costly to study owing to the effort and risk associated with accessing alpine areas, yet may be frequently spotted by mountain climbers who venture into alpine areas during their recreational pursuits [Jackson et al., 2015]. Citizen science is also used in fast-moving situations like natural disaster and to monitor long-term trends in the environment. For example, the British Columbia Big Tree Registry collates citizen science data on the locations of the largest trees in BC, thereby engaging citizens to help support policies to protect the largest trees in BC.

A useful starting point to check for citizen science datasets is the online data hub called SciStarter, which can be searched by keyword or location to identify citizen science projects around the globe. These datasets may be readily downloaded or downloaded through contacting the project leaders.

## Your Turn!

If you were to start a citizen science project to capture environmental data to inform public policy, what kind of information would you try to capture?

## 4.10 International Data

Some research questions extend beyond borders. For example, oceans are primarily international and data on oceans can be searched through the Ocean Biodiversity Information System. A database on food production and timber is published by the United Nations Food and Agricultural Organization. Academic research that attempts to answer environmental problems at the global scale now often publishes their datasets for open use, such as the global tree canopy height map [Potapov et al., 2021].

## 4.11 Metadata

To be written by Evan: at cursory level on how to collect and find metadata, why its important, how it can lead to other findings.

## 4.12 Unpublished Data and the Data Request

Governments manage a wide variety of data, which is sometimes located in relatively siloed ministries and departments. Datasets that are not readily accessible online, may still exist and can potentially be retrieved through a data request to the appropriate government agency. In the spirit of open data, many governments are becoming increasingly responsive to data requests, but success of this approach often hinges on connecting with the right person that may be able to help you.

Accessing data that are not readily available adds extra challenge but can reward you with new research and networking opportunities that can be highly beneficial for both parties. The data provider may benefit from the knowledge gained from your proposed research. They may be able to assist you with understanding the data, disseminating the final report, and even connecting you with job opportunities and other ways to continue your skill development. When sending a data request or data query, always be respectful of their time, and be tactful. A data request template is provided below:

1. Dear ... (person, or institution)
2. State your name and affiliation (e.g., University department and program/supervisor)
3. Briefly state your intended research or research aspiration (1-2 sentences)
4. State your **data inquiry** (e.g., do you know if x data exists?) or **data request** in bold text. Although you may not know exactly what you are

looking for, try to be as specific as possible on the type of data you are requesting by describing. Give your geographical area of interest if known either descriptively, in a map, or as a shapefile.

5. Thank them for considering your request.
6. If you do not hear back from them within 1-2 weeks reply back with another, much shorter email (e.g., *I'd like to follow up and ask if someone in your office may be able to respond to the above data request?*)

## 4.13 Historical Data Collections

**Historical data** collections generally include any spatial data source excluding satellite-based remote sensing that was produced prior to the widespread implementation of GIS in the mid 1990's. Historical data are typically not available as ready-to-use digital layers, and thus work is required up front to digitize them in preparation for spatial analysis.

Historical datasets can be extremely valuable in environmental research because they extend our ability to observe how the environment has changed over longer time horizons, potentially revealing vastly different landscapes and environmental conditions from those seen today. This insight can help remind us of levels of degradation or abundance that have become 'forgotten' by today's environmental managers, and can lead to surprising discoveries [McClenachan et al., 2015].

Although **historical datasets** can be very useful, they were often not collected for the intended purpose of being analyzed by future researchers. Data were often collected to serve the needs of the day, and were collected in a cost effective manner using tools and science that were available at that time. While this is less an issue for Census data, which has in some cases used relatively consistent survey questions through time, it complicates use of other datasets such as historical forest inventories, which have evolved their methods in step with technology and changing perceptions of how the forest ought to be monitored and valued. Thus, knowledge of how **historical data** were collected is sometimes required to accurately understand and interpret it. Overall, the process of locating, digitizing, and interpreting historical data can be a substantial portion of the work in a historical spatial analysis. In this section we cover historical **aerial photograph** collections, historical **natural resource administrative data** as well as **historical maps**.

## 4.14 Historical Aerial Photographs

The advent of **aerial photographs**, which are photographs of the Earth's surface taken from above (generally from an airplane), greatly improved mapping beginning in the 1930's and became the primary source of data for mapping land cover, timber volumes, topography, and national defense planning. Today, they

offer a valuable tool for the unique spatial and temporal resolutions they offer. Temporally, **aerial photos** offer snapshots of landscapes that predate satellite-based remotely-sensed data by many decades [Morgan et al., 2017], which can help inform restoration targets and cumulative effects assessments [Harker et al., 2021]. Aerial photos vary in their spatial resolution, but sometimes offer a surprisingly high spatial resolution that can be used to study fine-scale landscape attributes and their changes, such as stream courses [Little et al., 2013], fish habitat [Tomlinson et al., 2011], and soil hydrodynamics [Harker et al., 2021].

Using **aerial photographs** to track landscape change often requires first ‘tying’ them to the Earth to produce an orthophoto, a process discussed as it applies generally to image processing in Chapter 13 and discussed briefly here. An orthoimage is an aerial photograph or satellite imagery geometrically corrected so that the scale is uniform, such as in Figure 4.4. Unlike orthoimages, the scale of ordinary aerial images varies across the image, due to the changing elevation of the terrain surface (among other things). The process of creating an orthoimage from an ordinary aerial image is called orthorectification. Photogrammetrists are the professionals who specialize in creating orthorectified aerial imagery, and in compiling geometrically-accurate vector data from aerial images.

Compare the map and photograph below. Both show the same gas pipeline, which passes through hilly terrain. Note the deformation of the pipeline route in the photo relative to the shape of the route on the topographic map. Only the topographic map is accurate here. The deformation in the photo is caused by relief displacement. The photo would not serve well on its own as a source for topographic mapping.

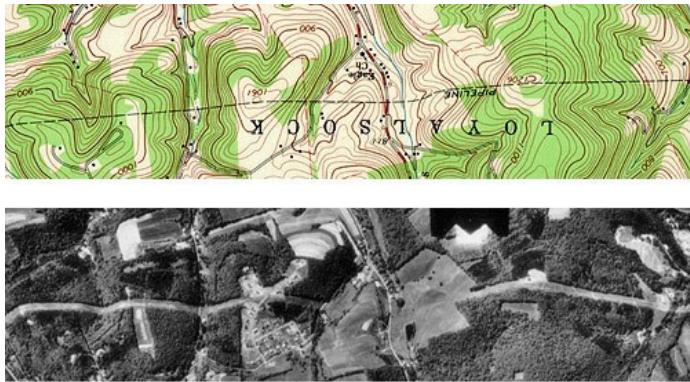


Figure 4.4: Example of how a linear feature can appear crooked in an aerial photograph that has not yet been orthorectified due to relief displacement. DiBiase [2014], CC BY 4.0.

Even in their un-orthorectified state, historical aerial photos can offer a powerful communication tool. They offer a window into historical landscapes that can be easily discerned and appreciated by viewers. Thus, even without orthorec-

tification and performing spatial analysis, historical aerial photos can enrich a research report and other communications.

## 4.15 Accessing Historical Aerial Photograph Collections

**Aerial photography** missions involved capturing sequences of overlapping images along parallel flight paths. A flight path produces a ‘roll’ of numerous adjacent images that overlap. Flight paths tend to be here and there, but not necessarily exactly where you need them! Therefore, the first step is to determine the availability of historical photographs rolls for your time frame and **area of interest**. Some collections can be searched relatively easily using a web-based GIS. For example, the Canada National Air Photo Library has a collection of roughly 6 million aerial photos some dating back to the 1920’s, which can be searched using the Earth Observation Data Management System. A search generally follows these steps:

1. Determine your area of interest.
2. Decide on the timeframe of interest.
3. Search via a GIS web map or paper flight line maps and examine which flight rolls cross over your time frame and area of interest.

Figure 4.5 shows the results from an example search. In this example, the area of interest (large pink rectangle, figure 4.5 was set by navigating to the study site within the web map then setting the current extent as the area of interest. Here the extent is centered on the coastline between St. John’s, Newfoundland and Cape Spear, the most easterly point in North America. We then searched for \*\*aerial photographs at three different time frames: 1940-1945 (figure 4.5) A), 1950-1955 (Figure 4.5) B), and 1960-1965 (figure 4.5) C). Indeed, aerial photos were found to be available at each period. The photos with smaller boxes (or foot prints) tend to have higher spatial resolution but cover less area. Assuming that fine spatial resolution is desired, the smallest photos have been selected in this example and could then be requested from the Library. Previews are often not available so we will not fully know the quality of the photos until we inspect them.

### Your Turn!

Go to the Canada Earth Observation Data Management System and search for historical aerial photos in your chosen area of interest using the time frames 1935-1950 and then 1950-1980. What is the oldest photo available?

If you searched but did not find anything helpful, don’t be discouraged. The area of interest in the example of Cape Spear, Newfoundland, happens to be a strategic location for national defense so it not surprising that it has excellent

#### 4.15. ACCESSING HISTORICAL AERIAL PHOTOGRAPH COLLECTIONS 103

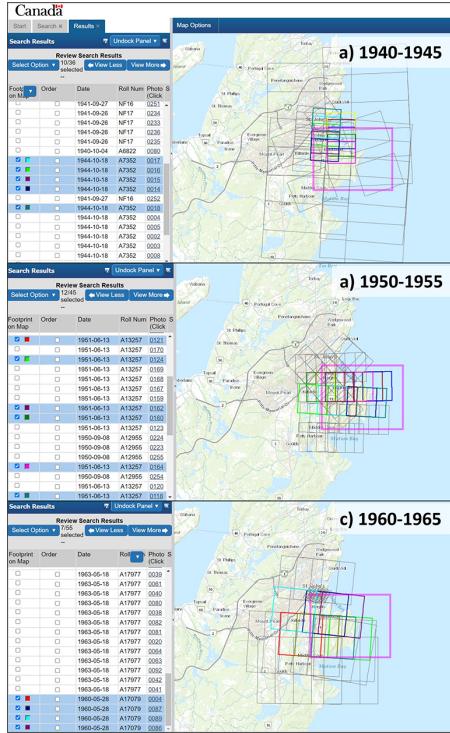


Figure 4.5: Example showing the availability of historical aerial photos in eastern Newfoundland at three time steps. Data from Natural Resources Canada [a], Open Government License - Canada.

coverage in the National Air Photo Library. In contrast, if you are interested in seeing an environmental feature such as historical forest cover in northern BC, recall that natural resources fall under the jurisdiction of provinces in Canada. Consequently, provinces may house aerial photo collections for your area. Some of these collections have been preserved by government or other institutions, such as the Geographic Information Center (GIC) at the University of British Columbia, which rescued a collection of 2.5 million aerial photos. These photos are available for researchers and commercial use. The GIC also maintains a list of other aerial photograph libraries, including for Alberta, Yukon, and the United States.

## 4.16 Natural Resource Administrative Data

Governments often conduct ecological and economic monitoring in their efforts to inform public policy and environmental management. Herein, this data are referred collectively to as natural resource administrative data. This data includes information collected during the process of administering natural resources use, such as to calculate fees, royalties, and licensing payments that the resource users must pay to the government for the use of public natural resources. Administering natural resources also requires monitoring data to spatially allocate harvest quotas on resources such as fish, big game, and timber. As opposed to remotely sensed data, this type of data often describes the actual amounts of natural resources available or used, and sometimes the number of users, who those users are, and what types of dependency they may have on the resources (e.g., their levels of income).

These data often come in a form called spatial panel data. Spatial panel data describe time series associated with particular spatial units (e.g., cities, wildlife management units, timber harvesting areas). Using spatial panel data typically requires:

1. downloading (or digitizing, if necessary) the statistical data as a spreadsheet
2. downloading the spatial geometry file
3. Linking the two files using an attribute join (chapter 5).

An example of a marvelous and yet relatively easy to use natural resource administrative data record is the BC big Game Hunting Statistics, which documents the number of large game hunted in BC by species, by hunter type (BC resident vs. non-resident hunter), and the effort (# days) that went into the hunts. This data can be made spatial by performing an attribute join with the BC Wildlife Management Units Layer. Attribute joins are discussed in Chapter 5).

Many natural resource administrative records are in digital form back to about 1980. Before that data often only exists in archival documents and must be digitized. Libraries are actively digitizing important archives, such as government annual reports, which are a rich source for natural resource administrative data.

## 4.17 Historical Maps

People have collected spatial information and mapped the world since long before GIS or aerial photos existed. Efforts are underway to preserve and digitize historical maps, and some collections are readily accessible. For example, insurance maps are maps made by insurance companies who mapped buildings, industrial complexes, and neighbourhoods to administer insurance policies since the late 1800's (e.g., for BC). Forest cover mapping became common in the early to mid 1900's (though, the early maps rarely survived) to estimate timber volumes. Natural disturbance mapping also became widespread in the early 1900's and considerable work has already been done to digitize and turn those data into readily usable forms (e.g., for wildfire and insect disturbance in BC). Land surveys dating back to the mid 1850's have also been used to systematically map historical forest cover, land ownership, and linear features such as roads [Tomscha et al., 2016].

Geographers recognize that all maps are subjective and **historical maps** are thus sometimes studied to understand how historical landscapes were perceived by society, revealing potential social biases and political orientations of who commissioned or created the map. This treads into the social sciences and humanities disciplines, which can offer additional and important ways to understand land management challenges today. For example, historical geographers have studied the history of fur trapline mapping because it offers insight into how First Nations traditional territories were ascribed into a form of information that could fit with the worldview of colonial governments [Iceton, 2019]. Thus understanding the transcription of these areas into maps which happened a century ago may help inform the complex spatial problem of how First Nations rights and titles to their traditional territories can be addressed in treaty negotiations and reconciliation.

## 4.18 Georeferencing Historical Maps

Although many types of data seem to be georeferenced, other information must be first processed into a form that can be analyzed. This is especially true for any data captured prior to when Global Navigation Satellite Systems (GNSS) became commercially available in the 2000's. For example, decades and sometimes centuries of data exist in the form of herbaria, ship logs, and tree ring records that offer salient information on the spatial distribution of biodiversity and natural processes. This information cannot readily be brought into a GIS. The solution is **georeferencing**, which is a process to assign non-spatial information a spatial location (x and y coordinates) based on a coordinate system. Here we discuss georeferencing as it applies to historical maps. To supplement this section, general theory is provided about georeferencing aerial images in Chapter 13.

A common use case for georeferencing in landscape studies is when a historical

map must be brought into GIS and overlaid with other data. Imagine you have a paper map and you use a desktop scanning device to scan it and save it as a digital image - this map depicts a particular area on Earth but there is no way for your computer to where and how on Earth to place this map (figure 4.6). In order to solve this problem, it is necessary to assign it geographic coordinate information so that GIS software can correctly align it with other georeferenced data.

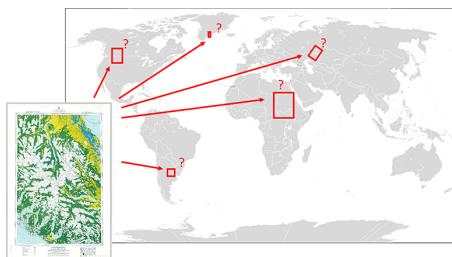


Figure 4.6: The need for georeferencing illustrated conceptually. Adapted from University of Texas Libraries [2021]. CC by 2.0.

Georeferencing is typically carried out using GIS software. The process of georeferencing varies slightly based on the GIS software you are using and the characteristics of the raster data you are working with, but the case study below provides a generalized workflow to help learn the overall process. Two important aspects are placing control points and rubbersheeting.

Control points are the locations on the map that we will use to tie our historical map into a coordinate system. Control points should be spaced evenly across the the map. There must be at least 3 control points, but preferably more (e.g., >10). Control points should be spaced relatively evenly to obtain a good rendering. Two options are discussed for control points

#### 4.18.1 Control Points on Maps with Grids or Graticule

Large area maps (e.g., an entire country or province) typically have graticule, which depict lines of latitude and longitude, and smaller scale maps often have UTM grids. These grids or graticule may span across the map, or just be located along the corner or edges of a map. Such maps can often be georeferenced in a GIS by first setting the desired coordinate system and then toggling on the grid or graticule within the GIS. Control points can be placed on the scanned raster at the line intersections than tied to the grid toggled on in the GIS. Here is a guide to **georeferencing** by map corners using QGIS.

#### 4.18.2 Grid and Graticule as Control Points

Not all maps have geographic coordinates on the map or along its corners (Figure 4.7 A). For such maps, control points must be placed on geographic features that



Figure 4.7: A comparison of A) a historical census map from 1931 with no graticule versus B) a 1961 census map with graticule representing latitude and longitude. Panel C) shows a close-up of the coordinate detail. Sutherland, CC BY 4.0.

can be linked to a base map that is already georeferenced and shows the locations of these features. Geographic features should be stable over time. For example, an ideal geographic feature is an island or cape in the ocean, or a mountain top. Be aware that many features do change over time: rivers meander, lakes are sometimes flooded by dam construction, and houses or other landmarks can be moved. In urban areas, try to identify features that have not changed over time. If using roads, use the center of road intersections.

### 4.18.3 Rubbersheeting

Once the control points are set, a transformation is applied to mold the historical map as best as possible into GIS space. The practice of using georeferencing historical maps using control points and transformations is an example of rubber sheeting. In cartography, rubbersheeting refers to the process by which a layer is distorted to allow it to be seamlessly joined to an adjacent geographic layer of matching imagery. This is sometimes referred to as image-to-vector conflation. Often this has to be done when layers created from adjacent map sheets are joined together. Rubber-sheeting is necessary because the imagery and the vector data will rarely match up correctly due to various reasons, such as the angle at which the image was taken, the curvature of the surface of the earth, minor movements in the imaging platform (such as a satellite or aircraft), and other errors in the imagery. A variety of transformations can be used during rubber sheeting. You should test a few to see how they work then choose one, which appears to produce the most satisfactory results in terms of the visual fit and lowest amount of error. If you are rubber sheeting multiple maps, it may be beneficial to use a consistent transformation to facilitate writing up your methods and communicating your research.

### 4.18.4 Documenting Georeferencing

During the process of georeferencing you must document the number of control points and the root mean square error (RMSE). Although there are multiple sources of uncertainty in the spatial precision of a historical map, uncertainty should be characterized where possible to demonstrate rigour in your methods and for communicating uncertainty.

## 4.19 Summary

Data are becoming increasingly accessible thanks to the open data movement, but one must still need to know where to find it. The search for data, whether social, environmental, or economic in nature, is facilitated by data repositories as well as informal approaches such as networking with colleagues, consulting data librarians, and reaching out to experts in your subject area. When data does not exist, we can sometimes create it. Historical data such as aerial photos, natural resource administrative data, and historical maps must often be digitized into a form useable for spatial analysis. However, this effort can be worth while for researchers interested in history and for the unique information gained on social and ecological change.

## 4.20 Reflection Questions

1. What are the key levels of Government where you live, and what kind of spatial data might each one manage?
2. What are two ways to find unpublished spatial data that is owned by a researcher?
3. What are the different types of data repositories where you can access spatial information?

## 4.21 Practice Questions

1. Try the case study on **georeferencing** a historical map. Record the number of control points placed, the RMSE, and the transformation use.
2. Draft a data request for a shapefile of bus routes as well as bus ridership statistics for the previous year in your hometown.

## 4.22 Summary

Data is becoming increasingly accessible thanks to the open data movement, but one must still need to know where to find it. The search for data, whether social, environmental, or economic in nature, is facilitated by data repositories as well as informal approaches such as networking with colleagues, consulting data librarians, and reaching out to experts in your subject area. When data does not exist, we can sometimes create it. Historical data such as aerial photos, natural resource administrative data, and historical maps must often be digitized into a form useable for spatial analysis. However, this effort can be worth while for researchers interested in history and for the unique information gained on social and ecological change.

GNSS and data transformations...

## Chapter 5

# Relational Databases

Written by Paul Pickell

You have almost certainly used a relational database in some form during your life, probably without even realizing it. Relational databases are foundational for information management in a GIS. In this chapter, we will look at the formal construction of relational databases, how they are used across a wide range of fields, and how we can use them to analyze spatial and aspatial information for environmental management.

## Learning Objectives

1. Identify the purpose of Relational Database Management Systems in GIS
2. Describe the elements of relational databases
3. Practice applying relational algebra and Boolean logic to relations
4. Recognize the uses of different keys for joining and relating information
5. Understand how to query relational databases in order to extract or produce new information

## Key Terms

Relational Database Management Systems, Tables, Relations, Rows, Tuples, Records, Columns, Attributes, Items, Structured Query Language, Boolean Logic, Relational Algebra, Entity-Relationship Model, Cartesian Product, Schema, Unary, Binary, Georelational Data Model, Domain, Symmetric Difference

## 5.1 Relational Database Management Systems

Suppose you have collected some data about some trees. You might have organized these data into a table, where each row represents a different plot, and each column represents some quantitative or qualitative measure about each record. How do you *manage* these data in order to extract useful information from your trees? This is where Relational Database Management Systems can help. A **Relational Database Management System (RDBMS)** is a software that allows the user to interact with tabular data. The basic services provided by a RDBMS include storing, querying, and manipulating relational databases. We say the databases are **relational** because they are based on a relational model first developed by Edgar Codd in the 1970s at IBM. The relational model for database management is distinguished from non-relational models by the fact that data are stored in highly structured tables instead of some other format like documents. This distinction is important, because the vast majority of GIS software utilize the relational model for database management.

## 5.2 Relational Databases

Within a RDBMS, we find **relational databases**, which are highly structured tables comprised of rows and columns. In fact, a table in a relational database is called a **relation**, a row is a **tuple**, and a column is an **attribute**. Relational databases are a great way to store simple data structures that can be organized into a relation with tuples and attributes. When we say that a table or relation is “structured”, we are referring to the fact that the data are organized according to a database **schema**, which is a set of constraints that ensure data integrity and consistency. For example, our set of trees likely all contain the same types of information and this can be easily organized into a relation. Suppose we measured the height, diameter at breast height (DBH), and species of each tree, then our relation would look like Figure 5.1.

As you can see from the example above, there are two components to geospatial data: the tabular data containing tuples and attributes and the spatial data that contain the coordinate pairs for a projected or geographic coordinate system. This structure is known generically as the **georelational data model**. Many formats of geospatial data conform to the georelational data model, which stores a relation of tuples and attributes separately from another relation containing the geometry and coordinates. These two tables are then dynamically related to one another in a RDBMS using GIS software. You will almost never interact or see the relation that stores the geometry and coordinates of features contained in a relational database. Instead, the GIS software manages those files in the background for the purpose of displaying a set of features on a map, and you primarily interact with the tabular data stored in the relation of tuples and attributes.

The schema for the very simple example above would include the constraint

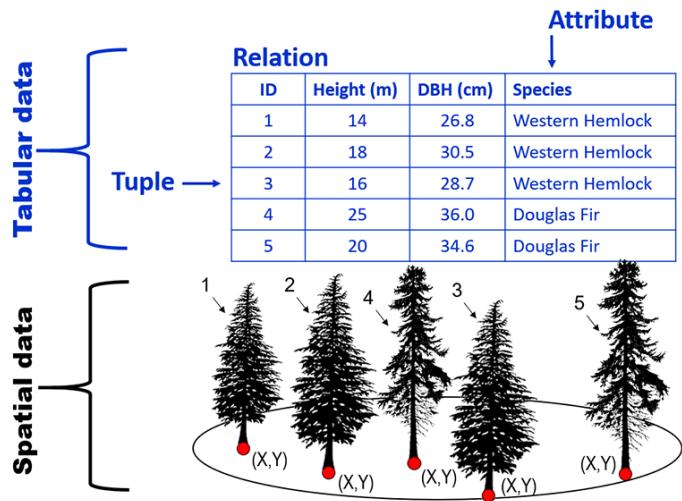


Figure 5.1: Tabular and spatial data are related by a Relational Database Management System (RDBMS) in a Geographic Information System (GIS). Images of Douglas-Fir and Western Hemlock trees by Natural Resources Canada [2013], Canadian Forest Service, modified with permission.

and expectation that when we retrieve the height of a particular tree from the relation, it will be returned to us as an integer number and not a date. This logic is extended to all attributes so that types of values are never mixed and values are never unexpectedly changed by any database operation. That is to say, we can and often do intentionally change values in a relation, but any new values must conform to the database schema for a particular attribute, which may also be constrained by a range and type of potential values, known as an attribute **domain**.

More formally, a relation  $R$  is a *subset* of two sets,  $A$  (tuples) and  $B$  (attributes). The product of these sets  $A \times B$  is called the **Cartesian product**. In the same way that Cartesian coordinates are ordered pairs of values from two axes, the Cartesian product of two sets gives us an ordered pair of elements  $(a, b)$  from sets  $A$  and  $B$ , where  $a$  is an element in the set  $A$ , written as  $a \in A$ , and  $b$  is an element of set  $B$ , written as  $b \in B$ . Therefore,  $R$  is both the Cartesian product as well as any subset of  $A \times B$ .

There are some important rules to be followed for organizing data into a relation:

1. Each tuple must share the same attributes as all the other tuples;
2. Each attribute has a unique name and is of the same *type* of data (i.e., integer, floating-point decimal, text, date, boolean, etc.);
3. The order of tuples and attributes can be rearranged without changing the meaning or integrity of the data;

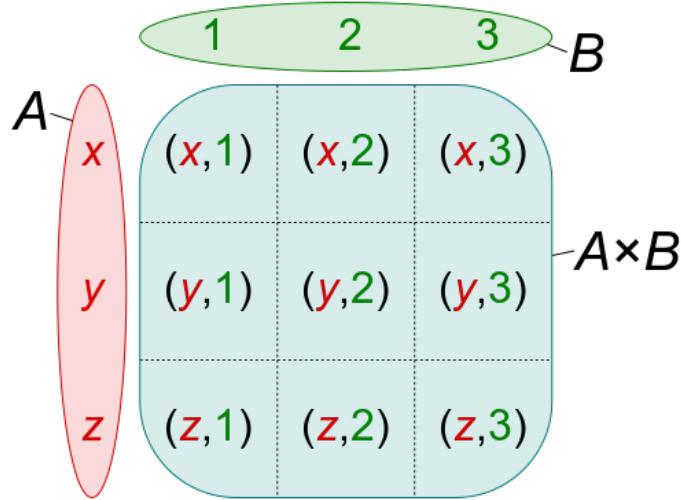


Figure 5.2: Cartesian product  $A \times B$  of A (tuples) and B (attributes). Image by Quartl [2012], CC BY-SA 3.0.

4. Each value of an element in a relation (i.e., combination of tuple and attribute) is *logically* accessible; and
5. Each tuple is unique (i.e., no duplicate observations).

If any of the above rules are broken, then  $R \neq A \times B$  and you are just looking at a plain-old table instead of a relation. In fact, Codd described a total of 13 rules for a RDBMS, but since this chapter is only a cursory introduction of RDBMS for GIS, you only need to be familiar with the five rules above. In this way, relational databases are comprised of relations that are highly structured by a schema, which allows the user to query, retrieve, update, and delete data using a RDBMS. At this point, you should understand that relational databases are highly structured so that we can apply logical expressions and languages to interact with the information contained within and between the relations. In the next two sections, we will look at how to apply two branches of mathematical logic to relations in order to extract useful information.

### 5.3 Relational Algebra

One of the fundamental jobs of a RDBMS is to apply relational algebra operations to relations stored in a relational database. Remember that we defined a relation as  $R = A \times B$  and that any subset of  $A \times B$  is also a relation. This transitive property of relations combined with the fact that relations are just sets allows us to apply set algebra. In other words, relational algebra operations take one relation as input and produce a new relation as an output without

modifying the input relation. This new output relation can then be used as an input to another operation because it is also a relation.

## 5.4 Selection

**Selection** is the simplest operation to understand and is probably the most-used in day-to-day GIS work. It does exactly what it sounds like, it retrieves a subset of a relation given some predicate or condition. For example, we could select all tree IDs from our relation  $R$  in Figure 5.1 that have a height greater than 20 m. This would yield tree ID=5. Formally, selection is expressed as  $\sigma_{\text{predicate}}(R)$  and the example above would be written as  $\sigma_{\text{height}>20}(R)$ , which evaluates to the following:

ID	Height (m)	DBH (cm)	Species
4	25	36	Douglas-Fir

## 5.5 Projection

If selection is understood to operate on attributes to return tuples, then **projection** is an operation on tuples to return attributes. For example, suppose we are only interested in the height and DBH attributes for the trees. We would use projection to return this new subset of the relation. Formally, projection is expressed as  $\Pi_{\text{predicate}}(R)$ . Both projection and selection are referred to as **unary** operators because they only require a single relation as input. The example above would be expressed using the attributes that we want to preserve, so  $\Pi_{\text{height}, \text{dbh}}(R)$ , which evaluates to the following:

Height (m)	DBH (cm)
14	26.8
18	30.5
16	28.7
25	36.0
20	34.6

At this point, it is important to emphasize the case of  $\Pi_{\text{species}}(R)$ , which evaluates to:

Species
Western Hemlock
Douglas-Fir

Recall that the output of a relational algebra operation is also a relation. Remember the rule that a relation cannot have any duplicate tuples? Well, in the case of a 1-dimensional relation where we only have one attribute and several tuples, any duplicate values for the tuples must be eliminated, leaving us with only

the two unique values “Douglas-Fir” and “Western Hemlock” when we project  $R$  over  $Species$ . You should recognize now that this property of projection can be useful for identifying the unique values of any attribute, which is frequently needed when sorting through a relational database.

## 5.6 Rename

**Rename** is an operator that allows us to assign a variable name to a relational algebra expression. This has the benefit of making it simpler to track or reuse previous operations in complex relational database algebra. For example, let  $S = \sigma_{height > 20}(R)$ , then  $\Pi_{species}(S)$  evaluates to:

Species
Douglas-Fir

## 5.7 Set Union

Next we will introduce **binary** operators, that is, they take two relations as input. **Set Union** is one such operator that effectively appends one relation to another. The important rule for union is that both input relations must share the same number and type of attributes or “union compatible”. Formally, set union is expressed as  $S \cup T$  where  $S$  and  $T$  are the two input relations. You can think of set union as simply concatenating the tuples of the two relations together. In other words, the tuples of  $S$  are appended to the tuples of  $T$  to generate a new output relation. For example, suppose that we make two subsets of our relation  $R$  of trees:

$$S = \sigma_{height < 20}(R)$$

$$T = \sigma_{height \geq 20}(R)$$

Then we can union these two relations back into our original relation  $R$  as  $S \cup T$ , which evaluates to:

ID	Height (m)	DBH (cm)	Species
1	14	26.8	Western Hemlock
2	18	30.5	Western Hemlock
3	16	28.7	Western Hemlock
4	25	36.0	Douglas-Fir
5	20	34.6	Douglas-Fir

Formally, this would all be expressed as  $\sigma_{height < 20}(R) \cup \sigma_{height \geq 20}(R)$  or  $S \cup T$ , which in this case is also just equivalent to  $R$ . You should see that the result of

union is an inclusion of all tuples, so semantically a union can be read as “the tuples in relation  $S$  or the tuples in relation  $T$ ”.

## 5.8 Set Intersection

On the other hand, suppose that we want to define a new relation based on restricting the set of tuples that are in two different relations. This is known as **set intersection** and is formally expressed as  $S \cap T$ . Just like union, intersection also requires that the two relations be union compatible. Suppose we have two relations defined by subsetting height by  $< 25$  m and  $> 15$  m:

$$S = \sigma_{height < 25}(R)$$

ID	Height (m)	DBH (cm)	Species
1	14	26.8	Western Hemlock
2	18	30.5	Western Hemlock
3	16	28.7	Western Hemlock
5	20	34.6	Douglas-Fir

$$T = \sigma_{height > 15}(R)$$

ID	Height (m)	DBH (cm)	Species
2	18	30.5	Western Hemlock
3	16	28.7	Western Hemlock
4	25	36.0	Douglas-Fir
5	20	34.6	Douglas-Fir

There are 3 tuples that appear in both of these relations, so the intersection  $S \cap T$  would evaluate to:

ID	Height (m)	DBH (cm)	Species
2	18	30.5	Western Hemlock
3	16	28.7	Western Hemlock
5	20	34.6	Douglas-Fir

Semantically, set intersection is read as “the tuples in relation  $S$  and the tuples in relation  $T$ ”.

## 5.9 Set Difference

**Set difference** returns the tuples that are unique in one relation relative to another relation, but both relations must be union compatible. Formally, difference is expressed as  $S - T$ , and just like mathematical subtraction, the order of

relations in the set difference is important and non-commutative. For example,  $\sigma_{height < 25}(R) - \sigma_{height > 15}(R)$  evaluates to:

ID	Height (m)	DBH (cm)	Species
1	14	26.8	Western Hemlock

and  $\sigma_{height > 15}(R) - \sigma_{height < 25}(R)$  evaluates to:

ID	Height (m)	DBH (cm)	Species
4	25	36	Douglas-Fir

Semantically, you would read the set difference  $S - T$  as “the tuples in relation  $S$  minus any of the same tuples in relation  $T$ ”.

## 5.10 Cartesian Product

So far, we have seen the cases of mathematical addition (set union) and subtraction (set difference), but we can also apply multiplication and division. Multiplication of two relations is simply known as the **Cartesian product**. In the same way that a set of tuples and attributes can be multiplied to create a relation  $R = A \times B$ , we can also multiply two relations together and they do not need to be union compatible. For example, if  $S = \Pi_{height, dbh}(\sigma_{height < 20}(R))$  evaluates to:

Height (m)	DBH (cm)
14	26.8
18	30.5
16	28.7

and  $T = \Pi_{ID, Species}(\sigma_{dbh > 34}(R))$  evaluates to:

ID	Species
4	Douglas-Fir
5	Douglas-Fir

then the Cartesian product of  $S \times T$  evaluates to:

ID	Height (m)	DBH (cm)	Species
4	14	26.8	Douglas-Fir
5	18	30.5	Douglas-Fir
4	16	28.7	Douglas-Fir
5	14	26.8	Douglas-Fir
4	18	30.5	Douglas-Fir
5	16	28.7	Douglas-Fir

## 5.11 Set Divison

Finally, **set division** is an operation of division between two relations, and you can think of it semantically as, “all the values of an attribute in  $R$  that are found with the tuples of  $S$ .” Set division is expressed as  $S \div T = U$  and like the Cartesian product and set difference, set division is non-commutative, so the order of  $S$  and  $T$  changes the value of  $U$ .

For the next example of set division, we will introduce a new relation  $S$ , which is not a subset of  $R$ . Suppose, in addition to  $R$ , we have cataloged information about different tree species, some of which are in  $R$  (these data are a small sample of a full list of tree species codes commonly used in British Columbia, Canada [Forest Practices Branch, 2005]):

Code	Species
AT	Trembling Aspen
BB	Balsam Fir
CW	Western Red Cedar
E	Birch
FD	Douglas-Fir
HW	Western Hemlock
YC	Yellow Cedar

Suppose we want to answer the question, *What are all the species codes that are present in our plot of trees?* We can answer this question by first projecting *Species* over  $R$  to give relation  $T = \Pi_{species}(R)$ :

Species
Western Hemlock
Douglas-Fir

Then, dividing  $S$  by  $T$ ,  $S \div T = U$ , can be formally expanded to:

$$\Pi_{code}(S) - \Pi_{code}((\Pi_{code}(S) \times T) - S)$$

We read the first term  $\Pi_{code}(S)$  as “the projection of the attributes of  $S$  that are not in  $T$ ”. In our case, there is only one attribute in  $S$  not in  $T$ , which is *Code*, so  $\Pi_{code}(S)$  evaluates to:

Code
AT
BB
CW
E
FD
HW
YC

Then,  $\Pi_{code}(S) \times T$  is the Cartesian product of the previous projection and  $T$ , which yields a relation of all the combinations of  $T$  with the attributes in  $S$  that are not in  $T$ :

Code	Species
AT	Western Hemlock
BB	Western Hemlock
CW	Western Hemlock
E	Western Hemlock
FD	Western Hemlock
HW	Western Hemlock
YC	Western Hemlock
AT	Douglas-Fir
BB	Douglas-Fir
CW	Douglas-Fir
E	Douglas-Fir
FD	Douglas-Fir
HW	Douglas-Fir
YC	Douglas-Fir

Next, we take the set difference between the Cartesian product above and  $S$ ,  $(\Pi_{code}(S) \times T) - S$ , which has the effect of removing the tuples already observed in  $S$ . This leaves us with a relation that has all the “incorrect” code-species combinations:

Code	Species
AT	Western Hemlock
BB	Western Hemlock
CW	Western Hemlock
E	Western Hemlock
FD	Western Hemlock
YC	Western Hemlock
AT	Douglas-Fir
BB	Douglas-Fir
CW	Douglas-Fir
E	Douglas-Fir
HW	Douglas-Fir
YC	Douglas-Fir

Next, we project  $Code$ , which again is the only attribute in  $S$  not in  $T$ , from the set difference above  $\Pi_{code}((\Pi_{code}(S) \times T) - S)$ , which yields:

Code
AT
BB
CW
E
FD
YC
HW

And finally, we take the set difference between  $\Pi_{code}(S)$  and the projection above to obtain the code for the trees in our plot:

Code
FD
HW

You can think of set division as the inverse of a Cartesian product. However, just like division, the Cartesian product itself is non-commutative because it is a set of *ordered* pairs. If  $S$  contains a tuple that is not in  $T$ , then the Cartesian product of  $S \times T$  has a different order than would be the case if both  $S$  and  $T$  were identical. As an example,  $S \times T$  evaluates to:

Code	Species
FD	Western Hemlock
HW	Western Hemlock
FD	Douglas-Fir
HW	Douglas-Fir

and  $T \times S$  evaluates to:

Species	Code
Western Hemlock	FD
Douglas-Fir	FD
Western Hemlock	HW
Douglas-Fir	HW

Therefore, we cannot simply rewrite  $S \div T = U$  as  $U \times T = S$ , but we could express  $U \div S = T$ , which evaluates to  $T$ :

Species
Western Hemlock
Douglas-Fir

We have now considered the eight primary relational algebra operators (selection, projection, rename, set union, set intersection, set difference, Cartesian product, and set division) that can be applied to relations in a RDBMS. In the next section, we will look at another set of logical operators known as Boolean algebra, which give rise to logical languages for interacting with a RDBMS.

## 5.12 Boolean Algebra

Whenever we create and solve an arithmetic or relational algebra expression, we usually focus on the *value* of the output. In other words,  $1 + 1$  evaluates to a value of 2. But we often need to evaluate the *truth* of a statement. For example,  $1 + 1 = 2$  is a *true* statement and  $1 + 1 = 1$  is a *false* statement. **Boolean algebra** seeks to express mathematical expressions in terms of *truth values*. Boolean truth values are usually expressed as *true* or *false*, but it is also common in computer programming languages and GIS to see these encoded with values of 1 for *true* and 0 for *false*. Attributes can also take on Boolean values of *true* or *false* as a data type. Boolean algebra uses equality and conditional operators, which we will consider next.

## 5.13 Equality Operators

You are probably already familiar with the basic equality operators used in Boolean algebra: - = “exactly equal to” (usually expressed with  $=$ ) -  $>$  “greater than” -  $\geq$  “greater than or equal to” (usually expressed with  $\geq$ ) -  $<$  “less than” -  $\leq$  “less than or equal to” (usually expressed with  $\leq$ ) -  $\neq$  “not equal to” (usually expressed with  $\neq$  or  $\neq$ )

All of the equality operators above evaluate to logical *true* or *false* values. They are quite elementary, so we will not go into much detail except to show that these equality operators are the basis for forming more complex Boolean expressions.

Basic arithmetic expressions can also be applied to Boolean truth values and it can be helpful to rewrite Boolean values with values of 1 and 0:

- $true + true = 2$
- $true + false = 1$
- $false + false = 0$
- $true - true = 0$
- $true \div false = undefined$
- $false \div true = 0$
- $true \times true = 1 = true$
- $5 \times false = 0 = false$

Multiplication of Boolean values is a special case where the expression will always result in a Boolean value. That is, multiplying any combination of 1 and 0 will always return 1 or 0. In other words, the domain of the input  $[0, 1]$  is equivalent to the domain of the output  $[0, 1]$ , which is a property that is frequently exploited in GIS in order to concatenate more complex expressions. For example, the statement  $true \times (1 + 2 = 3) \times (4 > 3)$  can be rewritten as  $1 \times 1 \times 1$  and evaluates to *true*, while  $true \times (1 + 2 = 3) \times (4 < 3)$  can be rewritten as  $1 \times 1 \times 0$  and evaluates to *false*.

Below are some examples of using equality operators and what they evaluate to:

$true = false$  can be rewritten as  $1 = 0$ , which is *false*.  $true > false$  can be rewritten as  $1 > 0$ , which is *true*.  $true \neq false$  is *true*.  $1 + 1 = 1$  is *false*.  $2 + 3 = 4 + 1$  is *true*.

Next, we will look at how to apply arithmetic and equality expressions to relations and evaluate them in Boolean terms. We have already seen how predicates allow us to do set selection with relational algebra. For example, we know that  $q$  evaluates to:

ID	Height (m)	DBH (cm)	Species
2	18	30.5	Western Hemlock
3	16	28.7	Western Hemlock
4	25	36.0	Douglas-Fir
5	20	34.6	Douglas-Fir

If we were to break this down in Boolean terms, the statement  $height > 15$  applied to  $R$  returns the following Boolean values for each tuple:

ID	Height (m)	DBH (cm)	Species	Boolean
1	14	26.8	Western Hemlock	FALSE
2	18	30.5	Western Hemlock	TRUE
3	16	28.7	Western Hemlock	TRUE
4	25	36.0	Douglas-Fir	TRUE
5	20	34.6	Douglas-Fir	TRUE

For another example, consider that  $(height > 15) \times (species = WesternHemlock)$  evaluates to:

ID	Height (m)	DBH (cm)	Species	Boolean
1	14	26.8	Western Hemlock	FALSE
2	18	30.5	Western Hemlock	TRUE
3	16	28.7	Western Hemlock	TRUE
4	25	36.0	Douglas-Fir	FALSE
5	20	34.6	Douglas-Fir	FALSE

We can also evaluate the equivalency between two attributes such as  $height = dbh$ , which evaluates to:

ID	Height (m)	DBH (cm)	Species	Boolean
1	14	26.8	Western Hemlock	FALSE
2	18	30.5	Western Hemlock	FALSE
3	16	28.7	Western Hemlock	FALSE
4	25	36.0	Douglas-Fir	FALSE
5	20	34.6	Douglas-Fir	FALSE

## 5.14 Conditional Operators

Now that we have a good understanding of equivalency operators, let us turn to consider conditional operators, which are also known as Boolean operators. Boolean operators are, in some ways, similar to some arithmetic operators except that they are based on natural language. There are three primary Boolean operators: *AND*, *OR*, *XOR*, and *NOT*. These operators are commonly used for database queries and with search engines, and indeed they form an important basis for query languages that are used to interact with an RDBMS.

Consider the statement  $(height > 15) AND (species = WesternHemlock)$ . This statement is equivalent to  $(height > 15) \times (species = WesternHemlock)$  and evaluates to exactly what we saw earlier:

ID	Height (m)	DBH (cm)	Species	Boolean
1	14	26.8	Western Hemlock	FALSE
2	18	30.5	Western Hemlock	TRUE
3	16	28.7	Western Hemlock	TRUE
4	25	36.0	Douglas-Fir	FALSE
5	20	34.6	Douglas-Fir	FALSE

Figure 5.3 illustrates what is going on here, we are only returning the tuples that evaluate *true* for both statements. Hence, Boolean *AND* is equivalent to multiplying two Boolean truth values together. You should also recognize that a Boolean *AND* is equivalent to what a set intersection  $A \cap B$  achieves between two relations.

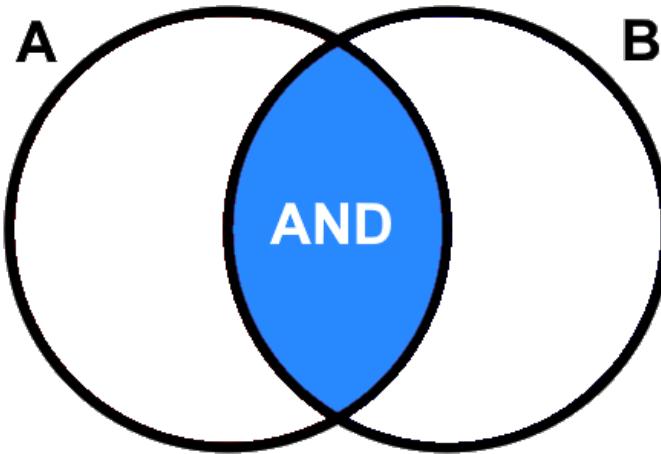


Figure 5.3: Boolean A AND B returns the area shaded blue. Pickell, CC-BY-SA-4.0.

If we do not want to be so restrictive, we could use Boolean *OR* such as  $(height > 15)OR(species = WesternHemlock)$ , which evaluates to:

ID	Height (m)	DBH (cm)	Species	Boolean
1	14	26.8	Western Hemlock	TRUE
2	18	30.5	Western Hemlock	TRUE
3	16	28.7	Western Hemlock	TRUE
4	25	36.0	Douglas-Fir	TRUE
5	20	34.6	Douglas-Fir	TRUE

Figure 5.4 illustrates the case of the Boolean *OR*. As you can see, it returns everything where either of the statements evaluate to *true*, regardless if the other statement is *false*. You should also recognize that a Boolean *OR* is equivalent to what a set union  $A \cup B$  achieves between two relations.

Suppose we want to identify all the trees that are greater than 15 m, but not Western Hemlock. In this case, we would use the expression  $(height > 15)NOT(species = WesternHemlock)$ , which evaluates to:

ID	Height (m)	DBH (cm)	Species	Boolean
1	14	26.8	Western Hemlock	FALSE
2	18	30.5	Western Hemlock	FALSE
3	16	28.7	Western Hemlock	FALSE
4	25	36.0	Douglas-Fir	TRUE
5	20	34.6	Douglas-Fir	TRUE

Figure 5.5 illustrates how Boolean *NOT* essentially negates or inverts the state-

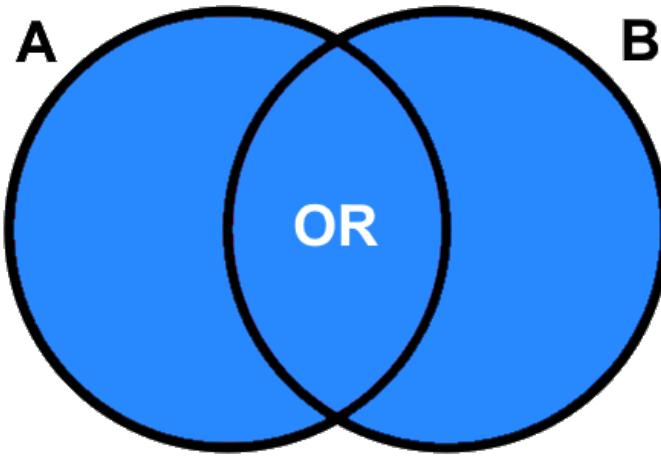


Figure 5.4: Boolean A OR B returns the area shaded blue. Pickell, CC-BY-SA-4.0.

ment that follows. In this case, instead of returning the Western Hemlock tuples,  $\text{NOT}(\text{species} = \text{WesternHemlock})$  returns “everything except” Western Hemlock, which is also equivalent to  $\text{species} \neq \text{WesternHemlock}$ .

Finally, the case of Boolean *XOR* returns any tuples that are *true* for both statements, but are *true* individually. This is known as the **symmetric difference** and “eXclusive OR” because we are only returning the tuples that are exclusive based on both statements. For example,  $(\text{height} > 15) \text{XOR} (\text{species} = \text{WesternHemlock})$  evaluates to:

ID	Height (m)	DBH (cm)	Species	Boolean
1	14	26.8	Western Hemlock	TRUE
2	18	30.5	Western Hemlock	FALSE
3	16	28.7	Western Hemlock	FALSE
4	25	36.0	Douglas-Fir	TRUE
5	20	34.6	Douglas-Fir	TRUE

Figure 5.6 illustrates how Boolean *XOR* excludes all the tuples that evaluate to *true* for both statements. In the example above, tuples ID=2 and ID=3 are excluded because both of the statements for height and species are *true*.

## 5.15 Joining Relations

More often than not, information is stored in separate relations, even if that information is about the same features like lakes, forests, or cities. Remember that a relation cannot have any duplicate tuples. This rule encourages the

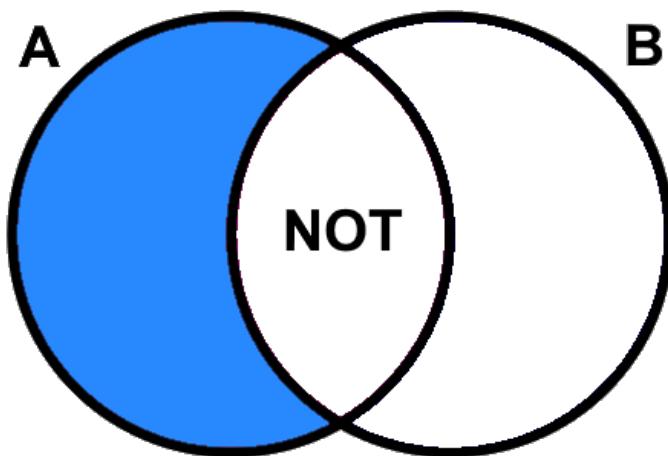


Figure 5.5: Boolean A NOT B returns the area shaded blue. Pickell, CC-BY-SA-4.0.

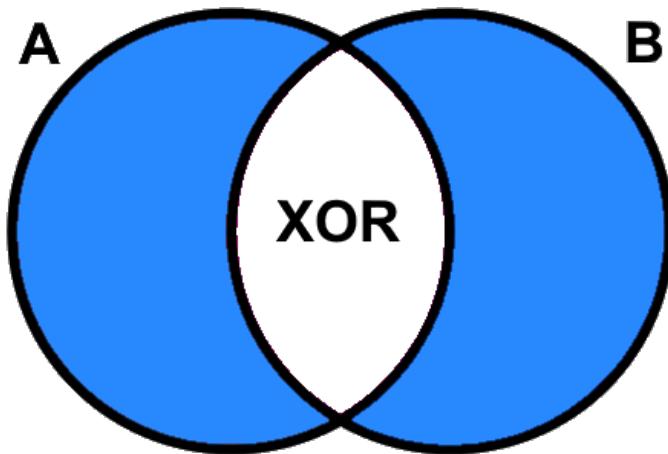


Figure 5.6: Boolean A XOR B returns the area shaded blue. Pickell, CC-BY-SA-4.0.

efficient storage and retrieval of information because information can be dynamically related as needed. For example, consider the overwhelming amount of information that is collected during a census. During the last census in 2016, there were over 14 million households in Canada. Can you imagine wielding a relation with 14 million tuples? These households can be segmented geographically by province, metropolitan areas, municipalities, and census subdivisions as well as by socioeconomic themes such as Indigenous peoples, age, sex, education, income, labour, housing, language, and others. Thus, those 14 million households can be divided up into many smaller relations, which can be accessed and summarized geographically and thematically. Since these relations represent different geographies or themes on the same set (i.e., households), we need to be more specific about how exactly two relations get combined if, for example, we want to combine themes with geographies. For this reason, we have joins.

## 5.16 Keys

Like the Cartesian product, joins are always binary operations, requiring two relations as input. While the Cartesian product combines relations by ordering all pairs of the elements from the two relations, we need a different method for correctly linking the tuples in relation  $R$  that correspond to the tuples of  $S$ . To do this, we rely on a common attribute called a **key**, which acts as an address between two relations. A **primary key** serves the purpose to identify the unique tuples in a relation and so it can be used to link other attribute information to those tuples. In a GIS, anytime that you create, copy or modify features such as points, lines or polygons, the newly created data layer (within the relational database) will be indexed with a primary key that counts from 1 to the number of features (tuples)  $n$  or from 0 to  $n - 1$ . For example,  $ID$  in our relation  $R$  serves as the primary key. There are other attributes in  $R$  that also uniquely identify all the tuples, but why do you think  $Height$  or  $DBH$  would be a poor operational choice as a primary key for a large field campaign?

While the primary key identifies the unique tuples in relation  $R$ , another key called the **foreign key**, serves to locate the same tuples in another relation  $S$ . In other words, a join is defined by a common attribute that is shared between two relations, the primary key in  $R$  and the foreign key in  $S$ . For example,  $Species$  is a foreign key in  $R$  and a primary key in  $S$ . The case of joining two relations using a set of attributes instead of a single attribute requires a **composite key**. For example, suppose we have a spatial dataset of all the municipalities across Canada. Some of these municipalities will share the same name, though they are in different provinces. Richmond is a city in British Columbia, Ontario, and Quebec. If we need to join census data to these spatial features, we would need to use a composite key comprised of  $CityName$  and  $ProvinceName$ .

## 5.17 Natural Join

A **natural join** restrictively joins two relations based on a set of common attributes. In this way, natural join is similar to a set intersection in that we are only combining tuples that share an attribute value and any tuples that do not share an attribute value in the other relation are dropped from the output. However, a natural join does not require that two relations be union compatible like a set intersection. Instead, the only requirement is that at least one attribute is shared between the two relations and has the same domain. Formally, natural join is expressed as  $R \bowtie S$ . and is sometimes referred to as an inner join. As an example, consider our example relations  $R$  and  $S$ :

$R$

ID	Height (m)	DBH (cm)	Species
1	14	26.8	Western Hemlock
2	18	30.5	Western Hemlock
3	16	28.7	Western Hemlock
4	25	36.0	Douglas-Fir
5	20	34.6	Douglas-Fir

$S$

Code	Species
AT	Trembling Aspen
BB	Balsam Fir
CW	Western Red Cedar
E	Birch
FD	Douglas-Fir
HW	Western Hemlock
YC	Yellow Cedar

The natural join  $R \bowtie S$  evaluates to:

```
## Joining, by = "Species"
```

ID	Height (m)	DBH (cm)	Species	Code
1	14	26.8	Western Hemlock	HW
2	18	30.5	Western Hemlock	HW
3	16	28.7	Western Hemlock	HW
4	25	36.0	Douglas-Fir	FD
5	20	34.6	Douglas-Fir	FD

## 5.18 Outer Join

An **outer join** joins all the tuples of two relations based on a common attribute. The result is similar to a set union, except the input relations do not need to be union compatible. Formally, an outer join or sometimes called a full join is expressed as  $R \bowtie S$ , which evaluates to:

```
## Joining, by = "Species"
```

ID	Height (m)	DBH (cm)	Species	Code
1	14	26.8	Western Hemlock	HW
2	18	30.5	Western Hemlock	HW
3	16	28.7	Western Hemlock	HW
4	25	36.0	Douglas-Fir	FD
5	20	34.6	Douglas-Fir	FD
NA	NA	NA	Trembling Aspen	AT
NA	NA	NA	Balsam Fir	BB
NA	NA	NA	Western Red Cedar	CW
NA	NA	NA	Birch	E
NA	NA	NA	Yellow Cedar	YC

## 5.19 Right and Left Outer Join

Sometimes, it may be desirable to join attributes or tuples from one relation, but not the other. For these cases, we can use either **right outer join** or **left outer join**. Formally, right outer join is expressed as  $R \bowtie S$  and evaluates to:

```
## Joining, by = "Species"
```

ID	Height (m)	DBH (cm)	Species	Code
1	14	26.8	Western Hemlock	HW
2	18	30.5	Western Hemlock	HW
3	16	28.7	Western Hemlock	HW
4	25	36.0	Douglas-Fir	FD
5	20	34.6	Douglas-Fir	FD
NA	NA	NA	Trembling Aspen	AT
NA	NA	NA	Balsam Fir	BB
NA	NA	NA	Western Red Cedar	CW
NA	NA	NA	Birch	E
NA	NA	NA	Yellow Cedar	YC

Formally, left outer join is expressed as  $R \bowtie S$  and evaluates to:

```
## Joining, by = "Species"
```

ID	Height (m)	DBH (cm)	Species	Code
1	14	26.8	Western Hemlock	HW
2	18	30.5	Western Hemlock	HW
3	16	28.7	Western Hemlock	HW
4	25	36.0	Douglas-Fir	FD
5	20	34.6	Douglas-Fir	FD

## 5.20 Theta Join

We can also join relations conditionally and without sharing a common attribute, which is known as a **theta join** and expressed as  $R \bowtie_{\theta} S$ . To understand how a theta join works, it is useful to recognize that  $R \bowtie_{\theta} S = \sigma_{\theta}(R \times S)$ . As you can see, a theta join is simply a selection of a Cartesian product where theta  $\theta$  is the predicate. For example,  $R \bowtie_{height > 19} S$  evaluates to:

```
## Joining, by = "Species"
```

ID	Height (m)	DBH (cm)	Species	Code
4	25	36.0	Douglas-Fir	FD
5	20	34.6	Douglas-Fir	FD

## 5.21 Cardinality of Joins

Depending on the schema of the two relations being joined, the number of tuples joined from one relation to another can vary and is known as **cardinality**. In the simplest case, one tuple in  $R$  is joined to one tuple in  $S$ , and this cardinality is known as **one-to-one** usually expressed as 1:1. The natural join example above,  $R \bowtie S$ , is an example of **one-to-many** (1:M) or **many-to-one** (M:1) cardinality because one species tuple found in  $S$  corresponds to many species tuples in  $R$ . Finally, **many-to-many** (M:M) cardinality describes the case where there are multiple tuples in  $R$  that correspond to multiple tuples in  $S$ . An example of a many-to-many relationship might be many species of trees in  $R$  that correspond to many forest stands in  $S$ . In other words, a forest stand might be comprised of many species and any particular species might be found in many forest stands. Figure 5.7 illustrates how cardinality might emerge depending on the relational schema and problem at hand.

## 5.22 Structured Query Language

Throughout this chapter, we have seen the various ways that relations are manipulated through relational algebra, Boolean logic, and joins. Since a GIS relies on a RDBMS to interact with data, especially data in the attribute table, geomatics professionals literally need a language to programmatically execute relational algebra, joins, and the other functions of a RDBMS within the GIS software.

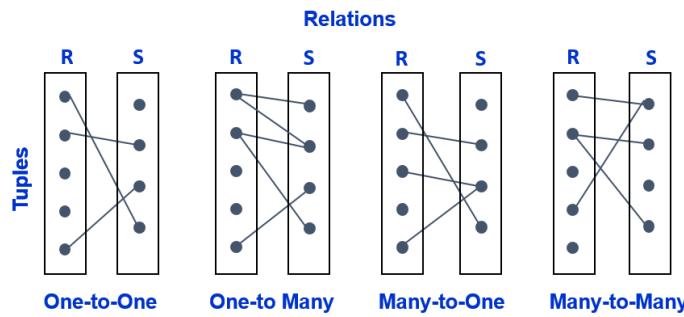


Figure 5.7: Cardinality of joins between relations R and S. Pickell, CC-BY-SA-4.0.

Such languages are known as query languages, each with its own syntax and use. By far, the most commonly used query language for RDBMS in GIS and across other systems is **Structured Query Language** abbreviated **SQL** and pronounced “sequel”. SQL has five primary language elements: 1. Clauses state an action or operation; 2. Expressions evaluate to a value; 3. Predicates evaluate an expression using equivalency and Boolean operators; 4. Queries apply set selection on a predicate; and 5. Statements are the combination of all the elements above

SQL has numerous keywords, which are the actions that comprise a clause. It is beyond the scope of this textbook to describe all of them, but most of the SQL keywords are implemented within GIS software in other ways. For example, you would rarely need to programmatically **ADD** an attribute to a relation. Instead, you might click an “Add field” button within the GIS software you are using. Similarly, you might never programmatically **UPDATE** the value for a tuple because most GIS software will allow you to simply double-click a cell in the table and change the value. The primary action that is nearly always performed programmatically with proper SQL syntax is applying a query, which is what we will focus on for this section.

SQL queries are fundamental for implementing set selection  $\sigma_{\text{predicate}}$  and they look like this:

```
SELECT attributes
FROM relation
WHERE predicate;
```

The entire form above is a statement, which is enclosed by a semi-colon at the end. The statement is comprised of three clauses using the keywords: **SELECT**, **FROM**, and **WHERE**. The **SELECT attributes** clause defines which attributes of the relation will be returned. You should recognize that this is the equivalent of applying a set projection  $\Pi_{\text{attributes}}$  to the entire set selection statement. You can specify attributes by name (e.g., **SELECT Species**), but it

is more common to return all of the attributes of the relation with an asterisk like `SELECT *`. The `FROM relation` clause defines which relation the selection is performed on. Keeping in mind that a RDBMS is comprised of many relations and at any given time you may have several different data sources open in your GIS software, the `FROM` keyword helps to clarify exactly which relation contains the attributes defined by the `SELECT` clause. Finally, the `WHERE predicate` clause defines the predicate that will be evaluated for the set selection, and this is where the magic happens. Although this is the formal syntax for a SQL query, most GIS software will usually only require the user to define the predicate, so next we will look at how to construct different SQL queries on our relation  $R$ .

Suppose we want to select the trees that are greater than 15 m, like in our previous equivalency example of  $\sigma_{height > 15}(R)$ . The SQL statement looks like this `SELECT * FROM R WHERE height > 15;`. If we only want to return the species for tree heights greater than 15 m, then the SQL statement looks like this `SELECT Species FROM R WHERE height > 15;` and evaluates to:

Species
Western Hemlock
Western Hemlock
Douglas-Fir
Douglas-Fir

The above SQL statement would be an example of  $\Pi_{species}(\sigma_{height > 15}(R))$ . In SQL, the multiplication symbol has the arithmetic meaning and cannot be used to concatenate two predicates. For this reason, we have the Boolean operators for evaluating multiple predicates. For example,  $(height > 15) \times (species = \text{WesternHemlock})$  would be written in SQL as `SELECT * FROM R WHERE height > 15 AND species="Western Hemlock";`. Our previous example of using Boolean *NOT* in SQL would be written as `SELECT * FROM R WHERE NOT species="Western Hemlock";`. These are all relatively simple examples, but it is common to create more complicated queries that use several Boolean operators. Note here how the species value in the expression above is in quotation marks "Western Hemlock" because the data type of the species attribute is a *string*. By contrast, the height value in the previous expression is simply an *integer number*. It is important to emphasize at this point that the only equivalency operator that can be used with string data type attributes is `=`. In other words, "Western Hemlock">>"Douglas-Fir" is illogical, cannot be evaluated, and will return an error.

If you combine two or more Boolean operators into one statement, then they are evaluated in SQL according to the following precedence: 1. Anything enclosed within parentheses () 2. NOT 3. AND 3. OR. For example, `SELECT * FROM R WHERE dbh < 30 AND species="Douglas-Fir" OR species="Western Hemlock";` would evaluate to:

ID	Height (m)	DBH (cm)	Species
1	14	26.8	Western Hemlock
2	18	30.5	Western Hemlock
3	16	28.7	Western Hemlock

But if we want the `OR` to be evaluated before the `AND`, then we need to use parentheses like `SELECT * FROM R WHERE dbh < 30 AND (species="Douglas-Fir" OR species="Western Hemlock")`;, which evaluates to:

ID	Height (m)	DBH (cm)	Species
1	14	26.8	Western Hemlock
3	16	28.7	Western Hemlock

You may notice the case of *XOR* conspicuously missing from the order above and this is because SQL does not natively implement the *XOR* operator. If you want to evaluate the exclusive OR example used in the previous section, (*height > 15*)*XOR(species = WesternHemlock)*, then you would construct a SQL statement like this `SELECT * FROM R WHERE (height > 15 OR species="Western Hemlock") AND NOT (species="Western Hemlock" AND height > 15)`;. As you can see, SQL queries can quickly get complex and involve many Boolean operators, so it is important to understand operator precedence and whenever in doubt, you can always use parentheses to override any precedence rules. You should also recognize that there are many ways to write complex statements to achieve your desired selection and you should always prefer the simplest statement possible.

Finally, a very common query that involves returning all tuples that match an attribute value in a list of values can be applied using the `IN` operator in SQL. For example, suppose we want to select all conifer tree species (codes: BB, CW, FD, HW, and YC) from *S* below:

Code	Species
AT	Trembling Aspen
BB	Balsam Fir
CW	Western Red Cedar
E	Birch
FD	Douglas-Fir
HW	Western Hemlock
YC	Yellow Cedar

Your natural reaction to this problem might be to write a long SQL statement like `SELECT * FROM S WHERE code="BB" OR code="CW" OR code="FD" OR code="HW" OR code="YC"`;. This is perfectly fine, but you can write this more economically with `IN` such as `SELECT * FROM S WHERE code IN("BB", "CW", "FD", "HW", "YC")`;. Be aware that a common mistake is to write a long predicate using `OR` like `code="BB" OR "CW" OR "FD" OR "HW"`

## 5.23. CASE STUDY: COMBINING SOCIOECONOMIC AND VEGETATION INFORMATION FOR ASSESSING POPULATION VULNERABILITY

OR "YC", but this is incorrect syntax in SQL. Remember that each side of an OR or AND operator is an *expression* that evaluates to a Boolean truth value. So `code="BB" OR "CW"` will return an error because "CW" alone cannot be evaluated to a Boolean truth value.

### 5.23 Case Study: Combining Socioeconomic and Vegetation Information for Assessing Population Vulnerability

*Case Study Author: Taelynn Lam (CC BY 4.0. unless otherwise indicated), University of British Columbia, Master of Geomatics for Environmental Management graduate, 2021*

Vegetation diversity in urban landscapes is important to support urban forest biodiversity and residents' mental health. The aim of this case study is to link together socioeconomic data and vegetation information to identify areas to prioritize intervention in the City of Vancouver. The Canadian Index of Multiple Deprivation (CIMD) [Statistics Canada, 2019] data has four dimensions of population vulnerability scores and we will aggregate these scores to obtain an overall vulnerability score for each dissemination area (DA) in Vancouver. We will compute the vegetation diversity score using street trees data [City of Vancouver, 2012] and vegetation type cover richness data<sup>1</sup> and then use query to identify priority areas.

### 5.24 Join

The raw CIMD tabular data includes the DA code and the corresponding vulnerability scores (table 1). In order to visualize these scores on a map, we will need to relate these scores to spatial data that include the information about the DA polygons and the coordinate pairs associated with each DA. Hence, a polygon shapefile of the DAs in Vancouver<sup>2</sup> is obtained, and its attributes are shown in Table 2.

The tabular data of the CIMD scores are related to the DAs polygon by the DA code. The cardinality of the relationship between these two tables is one-to-one as each DA is described by one set of the CIMD scores. To join the CIMD scores to the Vancouver DA polygons, we would use the PRCDDA attribute in the CIMD table as the foreign key to perform a join on the DAUID attribute in the Vancouver DA polygon relation. Now that the CIMD scores are joined to

---

<sup>1</sup>Obtained from reclassifying Land Cover Classification 2014 - 2m Raster to one vegetation class and five vegetation classes and counted the number of vegetation type cover classes using the Zonal Histogram Tool.

<sup>2</sup>Extracted by clipping the Canada-wide dissemination areas boundary to the City of Vancouver's municipality boundary.

Table 5.1: An excerpt of the CIMD data table.

PRCDDA	Province	DA population	Ethno-cultural composition quintiles	Ethno-cultur
59010123	British Columbia	434		1
59010124	British Columbia	559		1
59010125	British Columbia	522		2
59010126	British Columbia	671		2
59010127	British Columbia	319		1
59010128	British Columbia	545		3

Table 5.2: An excerpt of the Vancouver DA polygon shapefile attributes.

DAUID	
0	59150727
1	59150728
2	59150729
3	59150730
4	59150731
5	59150732

the Vancouver DA polygon attribute table, we can create choropleth maps to display the vulnerability scores of the DAs (Figure 5.8).

## 5.25 Calculation

Suppose we would like to calculate the overall vulnerability score for each DA. We would first name a new field (e.g., “aggregate\_score”), set the data type to double (to allow negative values and values with decimal places), and then enter the mathematical expression to specify the calculation to sum the four dimensions of CIMD scores and divide it by four to obtain an averaged vulnerability score for each DA. Using similar steps, we could apply a min-max normalization to transform this overall vulnerability score to a range between 0 and 1 to allow for a quick interpretation of the score. The formula is as follows:  $\frac{(X - X_{min})}{(X_{max} - X_{min})}$ .

Using what you have learned, join the street tree data and the vegetation type cover richness data to the Vancouver DA attribute table and to compute a vegetation diversity score. The street trees data shows the number of unique street tree species at a DA. Make sure you apply a min-max normalization to obtain the street tree diversity score. The vegetation diversity score can be computed by averaging the normalized scores of the two vegetation data.

Figure 5.9 shows a map of the vegetation diversity score at the DA level in

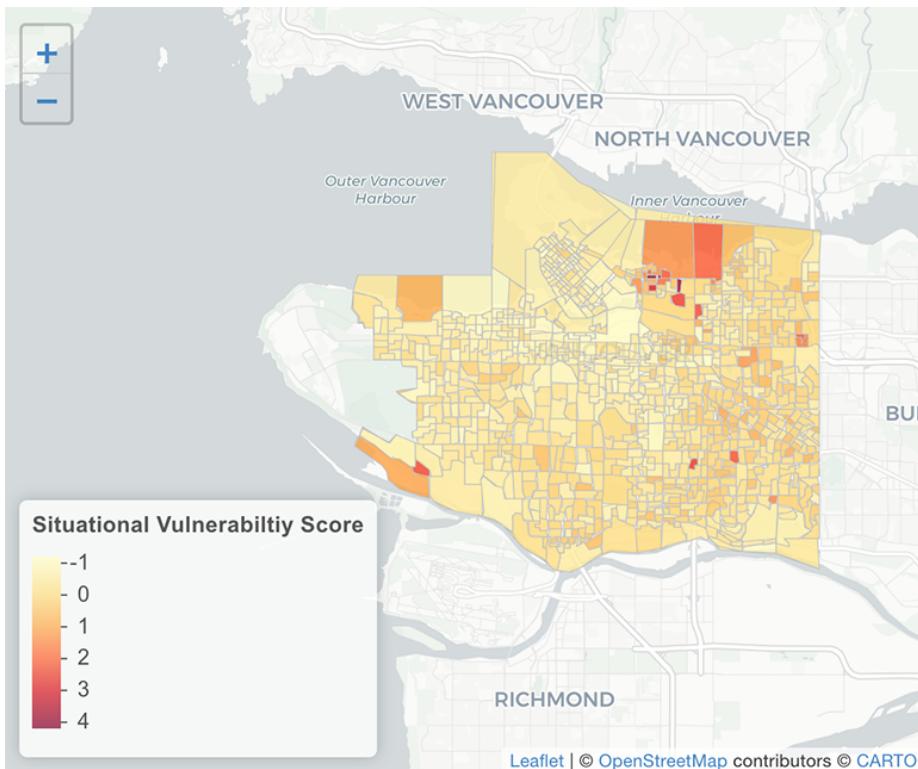


Figure 5.8: An example of choropleth map displaying the situational vulnerability scores in the City of Vancouver at DA level. Higher score represents the DA has higher situational vulnerable population e.g., population lacking a high school diploma, low-income population. Animated figure can be viewed in the web browser version of the textbook: <https://ubc-geomatics-textbook.github.io/geomatics-textbook relational-databases.html>

Table 5.3: An excerpt of the attribute table after the joins and calculations.

DAUID	Aggregated scores	Normalized aggregated scores	Species count	Street tree diversity	Vegeta
0	59150727	0.043	0.125	18	0.212
1	59150728	0.039	0.125	20	0.238
2	59150729	0.116	0.137	20	0.238
3	59150730	-0.103	0.101	24	0.288
4	59150731	-0.336	0.063	29	0.350
5	59150732	-0.095	0.103	33	0.400

Vancouver. The vegetation diversity score and the normalized aggregated vulnerability score are linked to each DA and can be viewed as you hover over the DA.

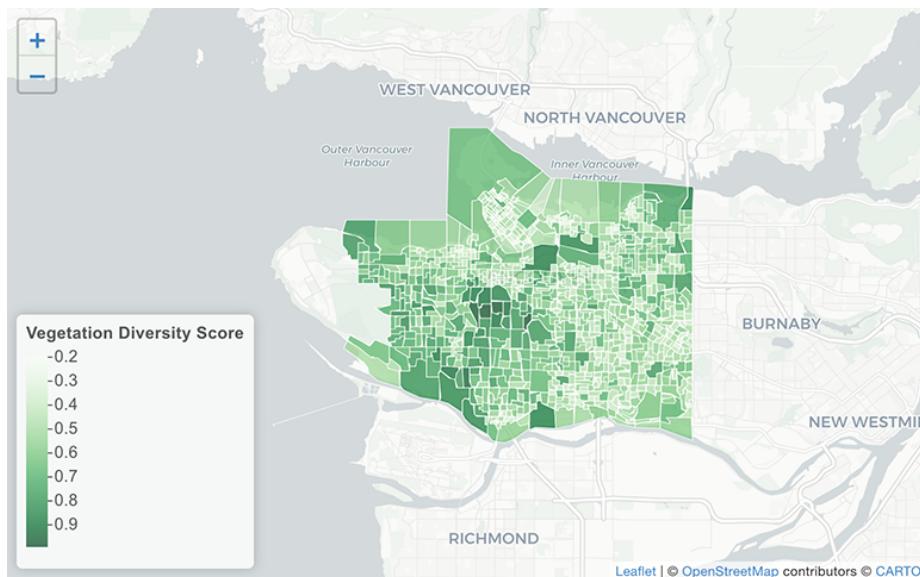


Figure 5.9: An example of choropleth map displaying vegetation diversity for each dissemination area in Vancouver. Data from City of Vancouver and licensed under the Open Government License - Vancouver. Lam, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://ubc-geomatics-textbook.github.io/geomatics-textbook/relational-databases.html>

## 5.26 Query

Areas with a higher proportion of vulnerable populations and less variety of vegetation to support resident's mental wellbeing are more in need for intervention. Supposed we define the priority area as DAs with a normalized aggregated vulnerability score greater than or equal to 0.5 and a vegetation diversity score less than 0.5. We could use the **Select By Attributes** tool to identify these priority areas by entering the appropriate query expression.

## Remember This?

Models are abstractions of reality and help us understand and communicate complex ideas.

Table 5.4: Query result shows five records matched the priority area requirements.

	DAUID	Normalized aggregated scores	Street tree diversity	Vegetation richness	Vegetation diversity
217	59150755	0.505	0.150	0.8	
222	59150760	0.590	0.075	0.6	
223	59150761	0.510	0.112	0.6	
527	59153187	0.522	0.125	0.8	
528	59153188	1.000	0.138	0.8	

## 5.27 Summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut in dolor nibh. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent et augue scelerisque, consectetur lorem eu, auctor lacus. Fusce metus leo, aliquet at velit eu, aliquam vehicula lacus. Donec libero mauris, pharetra sed tristique eu, gravida ac ex. Phasellus quis lectus lacus. Vivamus gravida eu nibh ac malesuada. Integer in libero pellentesque, tincidunt urna sed, feugiat risus. Sed at viverra magna. Sed sed neque sed purus malesuada auctor quis quis massa.

## Reflection Questions

1. Explain ipsum lorem.
2. Define ipsum lorem.
3. What is the role of ipsum lorem?
4. How does ipsum lorem work?

## Practice Questions

2. Given ipsum, solve for lorem.
3. Draw ipsum lorem.

## Recommended Readings

Ensure all inline citations are properly referenced here.



# **Chapter 6**

# **Overlay and Proximity Analysis**

Written by Amy Blood and Paul Pickell

Introduction here.

## **Learning Objectives**

1. Recognize the role of geoprocessing in applications of cartographic modeling
2. Understand the functions and opportunities of raster and vector overlay methods
3. Practice map algebra in raster overlay
4. Practice attribute transfer in vector overlay
5. Synthesize the role of relational databases in overlay analysis
- 6.

## **Key Terms**

Overlay, Union, Intersect, Identity, Difference, Symmetrical Difference, Buffer, Near Distance, Thiessen Polygons

## 6.1 Cartographic Modelling

### 6.1.1 Geoprocessing

### 6.1.2 Capability Modelling

The result of capability modelling is a binary classification of features or cells in a raster: 1 or 0; yes or no; true or false; capable or not capable. Recall from Chapter 3 that ordinal data scales are used to rank or order categorical or qualitative data elements. Even with just two classes, a capability map uses an ordinal data scale because 1 (capable) is better than 0 (not capable).

### 6.1.3 Suitability modelling

**Suitability modelling** is an extension of capability modelling that tells us how suitable a particular activity is for a given location. In other words, capability modelling gives us the spatial options that meet some minimum criteria and suitability modelling allows us to rank those options based on some attributes in our spatial data. From our earlier example, suppose we have identified 8 areas that are possible options for conserving habitat (i.e., capable), but we might only have the budget to proactively manage a few of these areas. So where should we prioritize our conservation and management activities? This is a question for suitability modelling!

Once we have calculated capability as an ordinal scale value (1 or 0), we can then use another set of attributes to calculate a **suitability score** for the capable areas. Frequently, the suitability score takes the form of a continuous ratio scale with values between [0, 1] or [0, 100] because we want to be able to place every capable feature on a spectrum from “least suitable” (0) to “most suitable” (1 or 100) based on some set of attributes. The calculation for the suitability score can take many forms and is dependent on the spatial problem that you are trying to solve. Some attributes can be used individually as a suitability score. For example, if bigger is better, then you could simply sort your capable features by area and you would have a suitability score on a continuous ratio scale, no further calculation needed. More commonly, we want to combine several attributes together in our scoring, which might represent data in different scales. Next, we will walk through an example for calculating a suitability score with nominal, ordinal, interval, and ratio data scales.

Suppose bigger really is better, so the area of the capable polygons will be one of our attributes for our suitability score, which is a ratio data scale. (By the way, you can extend this logic to lines and points as well: longer lines are preferred or a higher density of points is preferred.) Our first step here is to normalize these ratio data to a range of [0, 1] using the following equation:

$$X_{normalized} = (X - X_{min}) / (X_{max} - X_{min})$$

This is also called a min-max normalization, because the maximum value will be equal to 1 and the minimum value will be equal to 0:

ID	Area (ha)	Normalized Area (unitless)
1	9.96	1.0000000
2	7.02	0.6196636
3	6.46	0.5472186
4	6.15	0.5071151
5	5.00	0.3583441
6	3.33	0.1423027
7	2.80	0.0737387
8	2.23	0.0000000

Maybe our species is also found in several possible habitats. Habitat cover is a nominal data scale (e.g., “forest”, “wetland”, “non-forest”). If we know that our species is found in “forest” 60% of the time, in “wetland” 30% of the time, and in “non-forest” 10% of the time, then we can actually convert these nominal habitat covers into a ratio scale (i.e., 0.6, 0.3 and 0.1). In the case that you do not have additional numerical data to make this conversion, you could also make an educated guess and assign weights to your classes that sum to 1. For example, based on the literature, we might hypothesize that our species has preferences for “forest”, “wetland”, and “non-forest” that can be quantified with the weights 0.5, 0.25, and 0.25, respectively. Either approach is sensible, as long as you are transparent about your choice.

ID	Habitat (Nominal)	Habitat (Ratio)
1	Non-forest	0.1
2	Wetland	0.3
3	Forest	0.6
4	Non-forest	0.1
5	Forest	0.6
6	Wetland	0.3
7	Wetland	0.3
8	Forest	0.6

Maybe we also have land use intensities representing human activity and management that are classified as “high”, “medium”, and “low”. These land use intensities are an ordinal data scale. Frequently, ordinal data scales in geomatics are derived from some other numerical analysis. For example, land use intensity may have initially been mapped from the density of roads in an area or the frequency of a particular industrial activity, which was then classified into “high”, “medium”, and “low” intensity classes. It is worth looking at the documentation of the mapped data you are working with to see how the ordinal data were generated and what assumptions went into the terms used (i.e., “high”, “medium”, and “low”). For our example, let us assume that we know

nothing numerically about these land use intensities, just that “high” is worse for the conservation of our species than “low”. If we return to our original question, **where should we prioritize our conservation and management activities?**, then most likely our efforts will be best spent conserving areas that currently have high land use intensity. In this case, we can convert these ordinal data into ratio data by assigning weights that sum to 1. For example, “high” is 0.8, “medium” is 0.15, and “low” is 0.05.

ID	Land Use (Ordinal)	Land Use (Ratio)
1	High	0.80
2	Low	0.05
3	Medium	0.15
4	Medium	0.15
5	Medium	0.15
6	Low	0.05
7	High	0.80
8	High	0.80

Maybe we also have dates that represent the last year of a known disturbance like a fire. Dates are an interval data scale, but can easily be converted into a ratio scale by subtracting them from the current date to yield a measure of time-since something. For example, time-since last fire might be a good proxy for forage quality of our species regardless of the habitat cover. Once we have converted it to a ratio scale, we want to normalize it to a range of [0, 1], but suppose that more recent fire is better. In this case, we also need to reverse the range so that the oldest fire has a lower score than the most recent fire. We can achieve this by modifying the min-max equation so that we subtract  $X$  from  $X_{max}$ :

$$X_{normalized,reversed} = (X_{max} - X) / (X_{max} - X_{min})$$

ID	Year of fire (Interval)	Time-since last fire (Ratio)	Time-since last fire (Normalized)
1	1972	50	0.0000000
2	1975	47	0.0612245
3	1978	44	0.1224490
4	1982	40	0.2040816
5	1984	38	0.2448980
6	1999	23	0.5510204
7	2013	9	0.8367347
8	2021	1	1.0000000

Now we can put all the rescaled attributes together, add them up for each capable area, and divide by the total number of attributes that we used in our scoring process (four). This gives us an arithmetic mean that ranges between

$[0, 1]$  because all the other attributes also use this range. Sometimes this score is multiplied by 100 to convert the ratios into percentages. We can then sort the capable polygons in descending order by our suitability score:

ID	Area	Habitat	Land Use	Time-since last fire	Sum of Scores	Suitability
8	0.0000000	0.6	0.80	1.0000000	2.4000000	0.6000000
7	0.0737387	0.3	0.80	0.8367347	2.0104734	0.5026183
1	1.0000000	0.1	0.80	0.0000000	1.9000000	0.4750000
3	0.5472186	0.6	0.15	0.1224490	1.4196676	0.3549169
5	0.3583441	0.6	0.15	0.2448980	1.3532421	0.3383105
6	0.1423027	0.3	0.05	0.5510204	1.0433231	0.2608308
2	0.6196636	0.3	0.05	0.0612245	1.0308881	0.2577220
4	0.5071151	0.1	0.15	0.2040816	0.9611968	0.2402992

We can see from the above suitability analysis that capable polygon number 8 has the highest overall suitability. It is important to highlight two points here: the suitability score is unitless (after all, we have combined four very different data scales together); and the scores are on the same ratio scale, which means they can be directly compared. In other words, the most suitable location is more than twice as suitable as the least suitable location, based on the criteria and scoring scheme we used.

Note that the choice of weights for ordinal and nominal data scales are arbitrary, but these numbers can be based on an hypothesis, other numerical data, or your project's values. You can also iterate the suitability scoring process with different weights for ordinal and nominal scale data so that you achieve your desired project outcomes such as statistical distribution of scores or frequency of scores above a particular threshold value. For example, if you are trying to simultaneously solve the related question, **I have X dollars, how should I allocate them?**, then you might run a cost analysis that uses attributes accounting for the cost of intervening at a location. If you can convert your attributes into a dollar (ratio) scale, then you can simply add everything together to get the total cost for the activity at any given location.

For example, suppose that we know that our conservation intervention costs \$10,000 per hectare and we have a total budget of \$150,000 to conserve 15 total hectares. Looking at the capable areas in the earlier table, the total area that **could** be conserved is 44.22 hectares, which exceeds our budget. We need to solve two things: **where are our conservation efforts going to have the most impact on the species?** and **how can we allocate our budget efficiently to achieve that impact?** We have already solved the first problem, and the second problem is a matter of relating the costs to the suitability analysis, sorting the table based on the suitability score from our solution to the first problem, and then calculating a cumulative cost field that adds up the costs of the capable features in descending order of their suitability. This produces the following table:

ID	Area	Habitat	Land Use	Time-since last fire	Sum of Scores	Suitability	Cost (\$)
8	0.0000000	0.6	0.80	1.0000000	2.4000000	0.6000000	22300
7	0.0737387	0.3	0.80	0.8367347	2.0104734	0.5026183	28000
1	1.0000000	0.1	0.80	0.0000000	1.9000000	0.4750000	99600
3	0.5472186	0.6	0.15	0.1224490	1.4196676	0.3549169	64600
5	0.3583441	0.6	0.15	0.2448980	1.3532421	0.3383105	50000
6	0.1423027	0.3	0.05	0.5510204	1.0433231	0.2608308	33300
2	0.6196636	0.3	0.05	0.0612245	1.0308881	0.2577220	70200
4	0.5071151	0.1	0.15	0.2040816	0.9611968	0.2402992	61500

We can now see that prioritizing the top three suitable areas  $ID = 8, 7, 1$  for our conservation intervention will cost \$149,900, nearly exhausting our budget with \$100 left over for a party to celebrate the geomatics team. This is just one example of how cartographic modelling can provide powerful answers to very real spatial questions. Can you think of other mapped attributes, besides area, that could factor into this conservation cost analysis?

## 6.2 Overlay Methods

- 6.2.1 Attribute Transfer
- 6.2.2 Boolean Algebra
- 6.2.3 Spatial Join
- 6.2.4 Clip
- 6.2.5 Intersect
- 6.2.6 Line Intersection
- 6.2.7 Union
- 6.2.8 Identity
- 6.2.9 Erase
- 6.2.10 Split
- 6.2.11 Symmetrical Difference
- 6.2.12 Update

## 6.3 Proximity Methods

- 6.3.1 Euclidean Distance

Raster and vector

**6.3.2 Buffer****6.3.3 Attribute-Dependent Buffer****6.3.4 Near Distance****6.3.5 Thiessen Polygons****6.4 Summary**

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut in dolor nibh. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent et augue scelerisque, consectetur lorem eu, auctor lacus. Fusce metus leo, aliquet at velit eu, aliquam vehicula lacus. Donec libero mauris, pharetra sed tristique eu, gravida ac ex. Phasellus quis lectus lacus. Vivamus gravida eu nibh ac malesuada. Integer in libero pellentesque, tincidunt urna sed, feugiat risus. Sed at viverra magna. Sed sed neque sed purus malesuada auctor quis massa.

**Reflection Questions**

1. Explain ipsum lorem.
2. Define ipsum lorem.
3. What is the role of ipsum lorem?
4. How does ipsum lorem work?

**Practice Questions**

2. Given ipsum, solve for lorem.
3. Draw ipsum lorem.

**Recommended Readings**

Ensure all inline citations are properly referenced here.



# Chapter 7

# Topology and Geocoding

Written by Amy Blood and Paul Pickell

Frequently, we need spatial data to behave and relate in specific and predictable ways. Many types of analyses may expect spatial data to be represented and interact in a standard form. In this chapter, we will extend our knowledge of data models using topology, which unlocks many advanced spatial analyses. We will look at a specific example of an analysis that requires topology, geocoding, which will be a convenient segue into network analysis discussed in the following chapter.

## Learning Objectives

1. Understand the role of topology in governing data behaviour and data organization
2. Recognize some examples and uses of 2D and 3D topologies
3. Understand the role of bounding a set of points from triangulation and convex hulls
4. Synthesize the process of geocoding
5. Practice geocoding addresses and reverse geocoding addresses to other coordinate systems

## Key Terms

Vertex, Node, Pseudonode, Dangle, Planar Topology, Non-Planar Topology, Geocoding, Adjacency, Overlap, Connect, Inside, Reverse Geocoding, Singlearpart, Multipart, Holes, Delaunay Triangulation, Thiessen Polygons, Voronoi Diagram, Centroid, Convex Hull, Convex Alpha Hull, Multipatch

## 7.1 Topology

**Topology** describes the relationships of spatial data. This is a very broad definition that encompasses the wide range of possible arrangements of spatial data in practice. If we drill down into this concept, topology is really what allows us undertake specific types of analysis that requires or expects spatial data to behave in a certain way. If you think about the feature geometries that we have at our disposal, then there are no fewer than nine combinations of how these geometries can interact as illustrated in Figure 7.1 below.

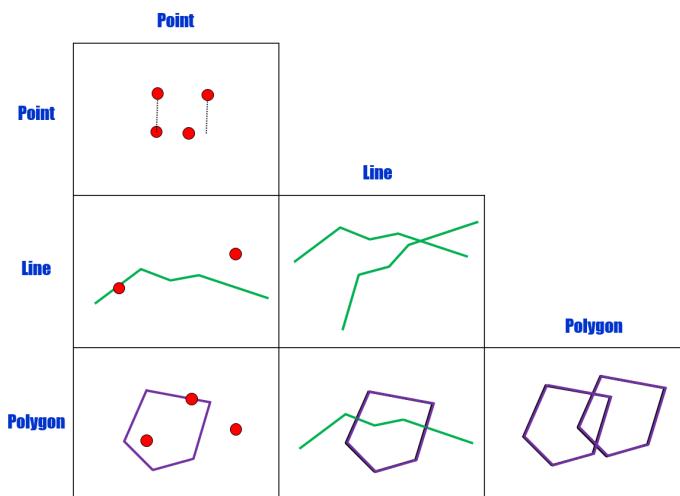


Figure 7.1: Grid showing all the combinations of how point, line and polygon geometries can interact. Pickell, CC-BY-SA-4.0.

It is important to recognize that there may be cases where we “expect” that a given combination of features will conform to a specific interaction. For example, the provinces and territories of Canada are typically represented as polygons that share adjacent boundaries. That is, the adjacent boundaries shared between any two provinces or territories cannot logically overlap as this representation (model) would contravene the legal definitions of the provinces and territories. In another case, a human technician may erroneously digitize a road that crosses another road without indicating that the two roads share an intersection, which could have consequences for how traffic flow can be modeled between the two roads (i.e., intersection with traffic light versus overpass). These are both examples of situations where topology is needed. Topology applies logic to define how features are expected to relate to other features in order to conform to knowledge systems like legal definitions of land and traffic flow. In short, topology ensures data integrity for other types of analysis.

## 7.2 Planar vs. Non-Planar Topology

In the context of topology, **planar** refers to the concept that all vertices of feature vector geometry are mapped onto the same plane. So in a planar worldview, all lines and polygons share coincident vertices. For example, if two polygons overlap, then the overlapping area forms a new polygon with a boundary of vertices defined by the union of the two other polygons. Also, if two lines overlap, then the two lines are divided into four new segments and a new vertex is formed at the intersection. In other words, planar topology does not allow polygons or lines to lay “underneath” or “on top” of another line or polygon and feature geometries must always be distinct.

On the other hand, **non-planar** topology is the concept that vertices of feature vector geometry can be mapped to different planes. It is important to emphasize here that when we are talking about planes that we are not referring to projected coordinate systems. It is generally assumed that any two spatial data layers containing feature geometries are interacting within the same projected coordinate system. Non-planar topology allows for other knowledge systems to be represented in spatial data. The case where a pipeline runs underneath a river or a territory that was traditionally used by several Indigenous peoples (Figure 7.2) are examples of valid non-planar topology.

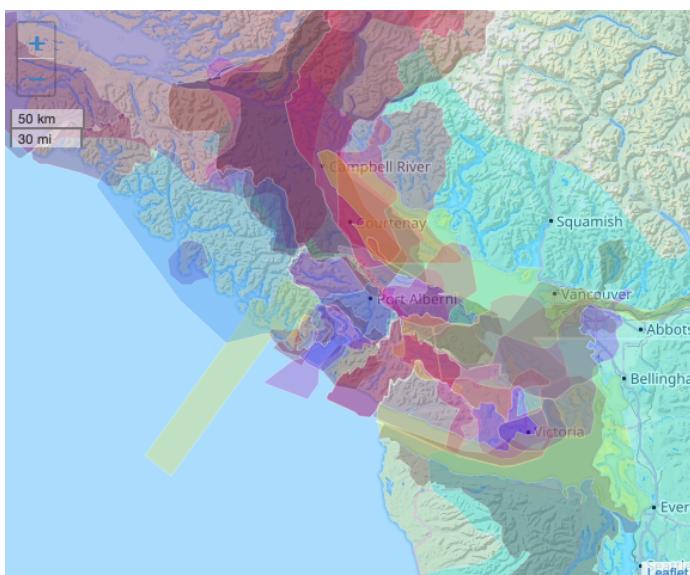


Figure 7.2: Non-planar topology of 36 indigenous territories overlapping Vancouver Island, British Columbia. Data from Native Land, CC0. Animated figure can be viewed in the web browser version of the textbook: <https://ubc-geomatics-textbook.github.io/geomatics-textbook/topology-and-geocoding.html>.

### 7.3 Implementing Planar Topology

Implementing planar topology involves defining specific rules for how features should relate to one another given some analytical context. This process also requires that the spatial data are housed a relational database or data model that supports topology. In other words, topology is enforced only by data models that support topological rules. When a topological rule is violated, the relational database identifies the contravening features and displays them on the map and in the attribute table. Then, it is up to an analyst to decide how the error should be corrected. For example, some errors like intersecting lines can automatically be split at the intersection while overlapping polygons might need to be manually edited to reflect the correct adjacency. Thus, the process of applying topology is first to work within a data model that supports topology, then choose the topological rules that reinforce a particular knowledge system, and finally to inspect and decide how to deal with any contraventions. Since planar topology is only supported by certain data models, and some data models are proprietary to certain software, the exact topological rules that can be implemented in a GIS are mostly dependent on the software that you are using. Instead of examining a specific GIS software package, we will discuss the “fundamental” planar topological relationships that are common across nearly all implementations of topology. (If you want to know more about how topology is implemented within specific data models, skip ahead to the “Data models supporting planar topology” section.)

So far, we have seen that there are six ways to combine feature geometries (points, lines, and polygons). We can extend this understanding to include at least six different ways that they can relate to one another: adjacent; overlap; intersect; connect; cover; and inside. Some of these relationships can be modeled *between* two different spatial layers (e.g., two point layers) or *within* a single spatial layer. In the following sections, we will look at different planar topological rules that apply both between and within feature geometries.

### 7.4 Adjacency and Overlap

There are times when we need to ensure that two polygons are **adjacent** to one another by sharing a common edge. If two polygons are not adjacent to one another, then a gap, known as a **sliver**, exists between them or they must **overlap**. Consider the case where we are mapping land covers. If we have a formal scheme that describes all possible land covers, then we expect that a map of land covers will have perfect adjacency between all polygons so that there are no areas that are not mapped (i.e., slivers) and that no area has multiple, overlapping land covers. Since lines are also 2-dimensional, lines can overlap other lines. Depending on the context, a topological rule may be needed to promote or prevent this relationship. For example, if you are modeling bus routes, then one road might support several different routes.

[figure of sliver] [figure of overlap]

Some examples of adjacency and overlap topological rules:

- Polygons within the same layer must not have gaps
- Polygons within the same layer must not overlap
- Polygons must not overlap other polygons
- Lines must not overlap other lines
- Lines must not self-overlap

## 7.5 Intersect and Connect

As we have seen from Chapter 3, lines are often used to represent phenomena that flow, so intersection and connection are important concepts for these representations. Important to understanding how connection and intersection work in planar topology, we need to understand that lines are comprised of a set of vertices and nodes. A **node** is simply the terminating vertex in a set of vertices for a line. For example, suppose the line segment  $A$  has a set of vertices,  $[[1,0], [1,3], [1,5]]$ . Then the nodes for  $A$  are  $[1,0]$  and  $[1,5]$  (Figure 7.3). Since nodes define the end points of a line segment, they are key to enforcing connection rules. We will look at network analysis in more detail in the next chapter. For now, let us consider two different networks that can help us conceptualize some fundamental line topology using nodes.

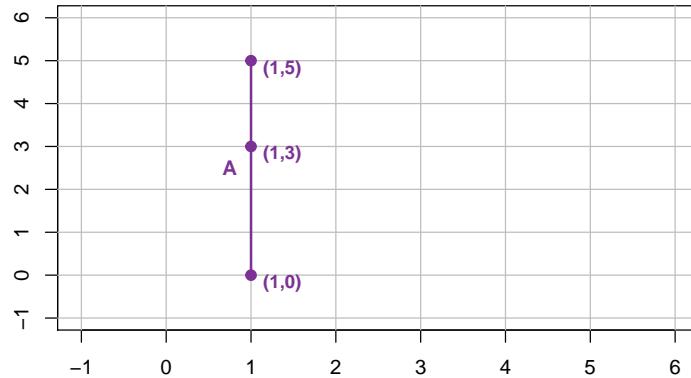


Figure 7.3: Lines are always comprised of two nodes. Line A shown here has nodes at  $[1,0]$  and  $[1,5]$ . Pickell, CC-BY-SA-4.0.

A network of streams and rivers is based on the hydrological knowledge system that explains how water moves over a terrain surface. In both theory and practice, we know that water flows from higher elevations to lower elevations with limited exceptions. Thus, we expect that streams will connect with other streams and continue to flow towards some outlet such as an ocean. **Connection** refers to the fact that the endpoint node of one stream will fall somewhere on another stream segment. Where two line segments come together, it is possible for one segment  $A$  to “undershoot” the other segment  $B$ , resulting in the end node of segment  $A$  appropriately named a **dangle** (Figure 7.4) and a loss

of connection.

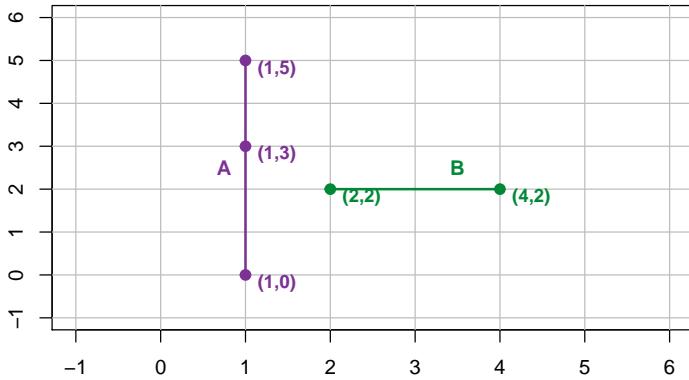


Figure 7.4: A dangle forms when a line (B) does not connect to another line (A). Pickell, CC-BY-SA-4.0.

Dangles are the opposite case to **intersections**, which occur when two line segments cross each other. With planar topology, intersections must be modeled with a shared node representing the intersection location. For example, suppose line segment  $B$  has a set of vertices,  $[[0, 1], [2, 1], [4, 1]]$ . If line segments  $A$  (defined above) and  $B$  are mapped together with non-planar topology, then they will intersect at  $[1, 2]$ , which is not a vertex represented in either segment (Figure 7.5).

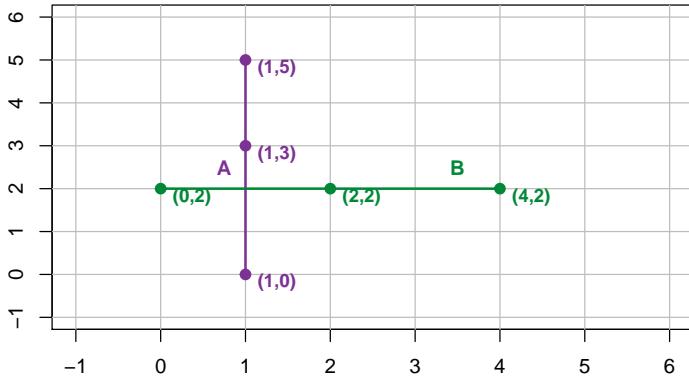


Figure 7.5: Line A mapped with Line B in non-planar topology. Pickell, CC-BY-SA-4.0.

Thus, the intersection of  $A$  and  $B$  with planar topology would yield four new segments:  $C = [[0, 2], [1, 2]]$ ,  $D = [[1, 2], [1, 3], [1, 5]]$ ,  $E = [[1, 2], [2, 2], [4, 2]]$ , and  $F = [[1, 0], [1, 2]]$ . Figure 7.6 illustrates how all four of these new segments share the same node  $[1, 2]$  at the intersection of  $A$  and  $B$ .

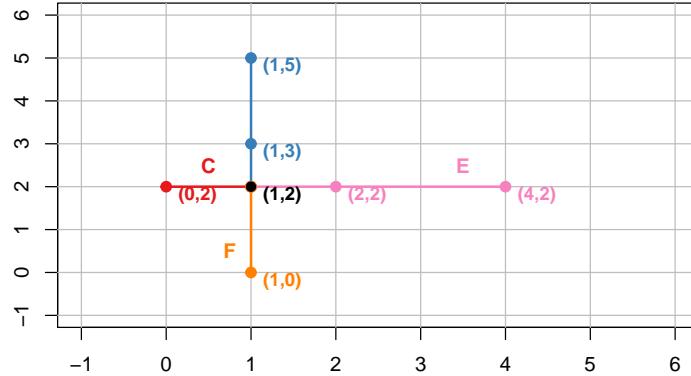


Figure 7.6: Line A mapped with Line B in planar topology yields segments C, D, E, and F. All segments share (1,2) as a node. Pickell, CC-BY-SA-4.0.

As well, **pseudonodes** can occur when a node does not actually terminate a line segment at a junction, for example, between two streams or roads. In other words, a pseudonode is a node that is shared by two lines. Figure 7.7 illustrates a pseudonode occurring at [3,5].

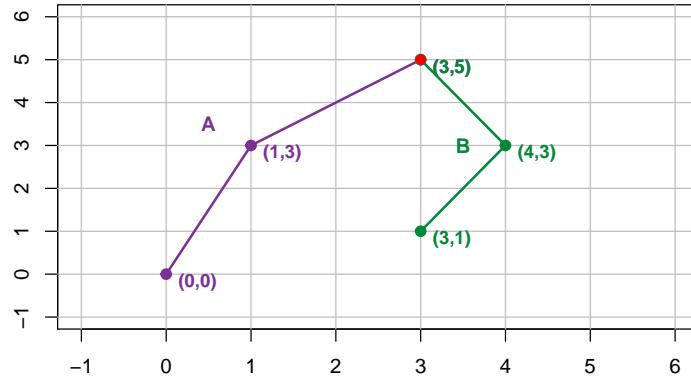


Figure 7.7: Lines A and B share a pseudonode at [3,5], indicated in red. Pickell, CC-BY-SA-4.0.

Some examples of intersection and connection topological rules:

- Lines must not intersect other lines
- Lines must intersect other lines
- Lines must not self-intersect
- Lines within a same layer must not self-intersect
- Lines must not have dangles

## 7.6 Coincident and Disjoint

Point features can be either **coincident** or **disjoint** with other point features. Point features that need to be disjoint may be representing trees, mountain peaks, or any similar type of feature that would be expected to be discrete in geographic space. There are also instances where we might need one set of point features to be coincident with another such as field plots that are centered using a tree or other spatially-discrete feature on the landscape.

Some examples of coincident and disjoint topological rules:

- Points must be disjoint with other points
- Points must be coincident with other points

## 7.7 Cover

**Cover** refers to planar topology where a feature lays on or within another feature. For example, dams represented as point features must be covered by a line representing a river (Figure 7.8). Similarly, lines representing rivers must be covered by polygons representing watersheds. As well, property parcel polygons must be covered by the municipal or regional tax authority polygon.

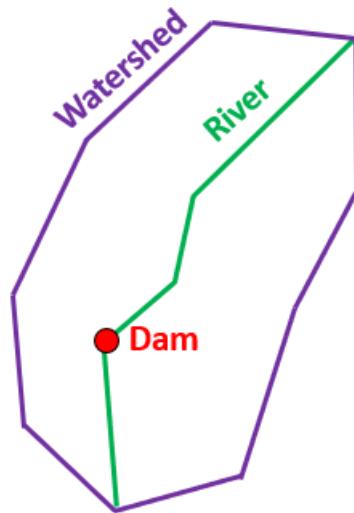


Figure 7.8: Topological relationship between dam (point) covered by a river (line), which is covered by a watershed (polygon). Pickell, CC-BY-SA-4.0.

Some examples of cover topological rules:

- Point must be covered by a line
- Point must be covered by a polygon
- Line must be covered by a polygon
- Polygon must be covered by a polygon

### 7.7.1 Multipart geometry

Sometimes we need to represent several points, lines or polygons as a collection, which is known as **multipart geometry**. Multipart geometry allow us to represent several disjoint and non-adjacent geometries as a single feature. In this way, we can assign attribute values to the collection of features rather than each geometry individually. The territorial boundary of Canada is a good example of an instance where a multipart geometry can be useful because all of the contiguous land and non-contiguous land (i.e., islands) can be represented and associated with a single feature in the attribute table. However, if the distinction of features is important, such as identifying the names of islands in the Haida Gwaii archipelago, then a singlepart geometry should be used (Figure 7.9).

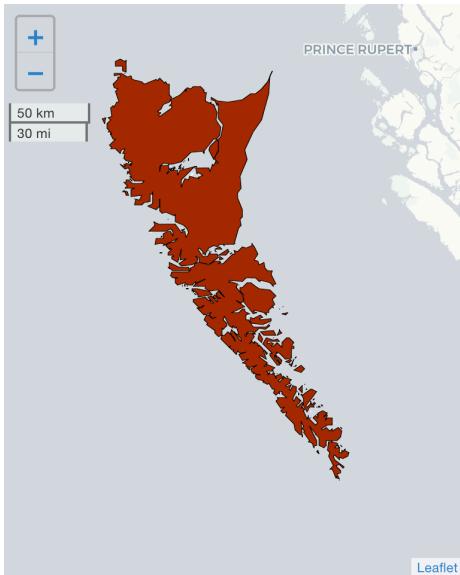


Figure 7.9: Singlepart geometry of the Haida Gwaii archipelago off the west coast of British Columbia, Canada. Hover over the islands to see the names. Polygon data from Statistics Canada and island placenames from Natural Resources Canada [b]. Open Government License - Canada. Animated figure can be viewed in the web browser version of the textbook: <https://ubc-geomatics-textbook.github.io/geomatics-textbook/topology-and-geocoding.html>

Although it is possible to convert from a multipart geometry into singlepart geometry, you need to carefully consider how your features should be represented in the attribute table. For example, if you will be undertaking calculations using area or perimeter of the constituent polygons that comprise a multipart geometry of Canada, then you will return a single value for all of Canada while singlepart geometry would return values for each individual polygon. As well,

area calculations can vary between multipart and singlepart geometry. For example, approximately 27% of Canada's land area (including freshwater), is comprised of more than 52,000 islands, which is a statistic you could only calculate with singlepart geometry. Thus, the choice of representing a feature using singlepart or multipart geometry should be based on how the features will be used in your analysis (i.e., aggregated versus disaggregated).

### 7.7.2 Holes

When dealing with polygon features, **holes** may occur, which represent discontinuity of the interior polygon space. Imagine the case of a forested land cover that surrounds a lake. If we consider the forested land cover polygon on its own, then the polygon will have a hole where the lake exists (Figure 7.10).



Figure 7.10: Conceptual forest land cover polygon that contains a lake causing a hole. Pickell, CC-BY-SA-4.0.

Topologically, holes in polygons imply that another polygon shares an adjacent boundary where the hole exists, for example, from the union of two layers (see Chapter 6). In our example, the lake would comprise its own polygon that would completely fill the hole.

### 7.7.3 Delaunay triangulation

**Delaunay triangulation** is method for forming a triangle mesh over a set of points. The Delaunay triangulation method connects all points in a set such that no point in the set lays *within* a circumcircle formed by any of the triangles in the mesh [Delaunay, 1934]. A circumcircle is a circle that passes through all the vertices of a cyclic polygon such as a triangle. In other words, the circumcircles are empty. To illustrate this, consider the four points in Figure 7.11. There are only two circumcircles that can be formed from this set of points that ensures that no point lays within a circumcircle. The triangulation is then simply the lines connecting the three points that fall on any given circumcircle. One important property of the Delaunay triangulation is that the smallest angle in the resulting triangles is maximized from the circumcircle fitting, which minimizes sliver triangles that might form with very shallow angles.

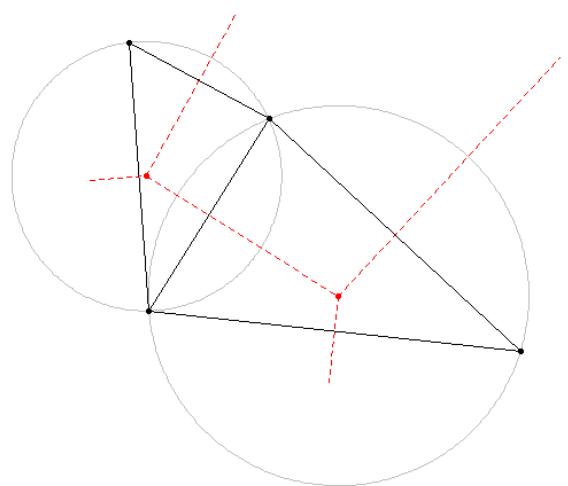


Figure 7.11: Delaunay triangulation of four points. Black lines show the triangulation, grey lines represent the circumcircles connecting the three points of each triangle, red points represent the centres of the circumcircles, and the red dotted lines show that connecting the centres of the circumcircles forms the Voronoi diagram. Pickell, CC-BY-SA-4.0.

Figure 7.12 shows a Delaunay triangulation for a set of 50 points. We can see that sliver triangles mostly occur on the edge of the extent of the points. Delaunay triangulations can be performed both in 2- and 3-dimensional Euclidean space and are therefore important for representing 3D surfaces as well as performing spatial estimation over 2D areas from a set of points.

#### 7.7.4 Multipatch geometries

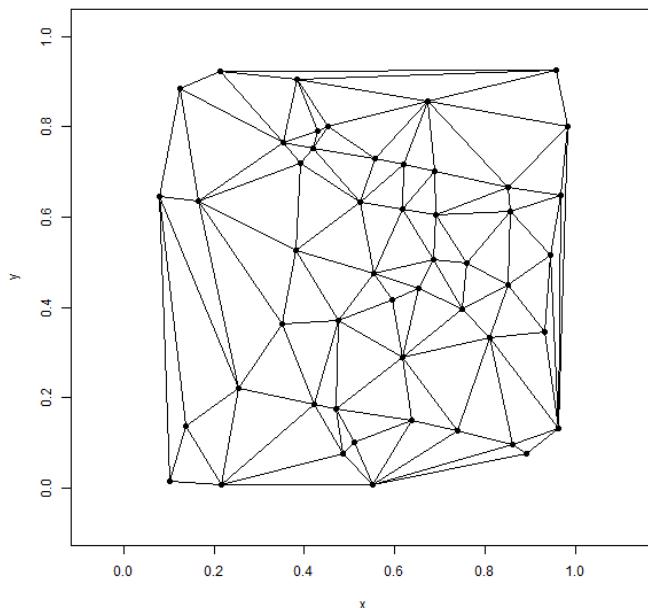


Figure 7.12: Delaunay triangulation of 50 random points. Pickell, CC-BY-SA-4.0.

#### 7.7.5 Thiessen polygons

**Thiessen polygons** are an implementation of a nearest neighbour algorithm in Euclidean space: given some set of input point features mapped on a plane, partition the plane into polygon areas that represent the nearest locations on the plane to those points. These resulting polygons are also sometimes referred to as proximal polygons, representing the proximal areas given some set of points. When Thiessen polygons are created for geographic data, the resulting diagrams are called **Voronoi diagrams** and sometimes referred to as Voronoi maps (Figure 7.13). Voronoi maps have many uses such as partitioning geographic space into areas that are nearest to weather stations, airports, or cellular towers. Thiessen polygons can be intersected with other geographic data layers

in a GIS using map algebra to efficiently solve proximal questions like, “what is the nearest X?” without having to search or calculate the exact distances of all nearby features, which can be computationally time-consuming [Okabe et al., 1994].

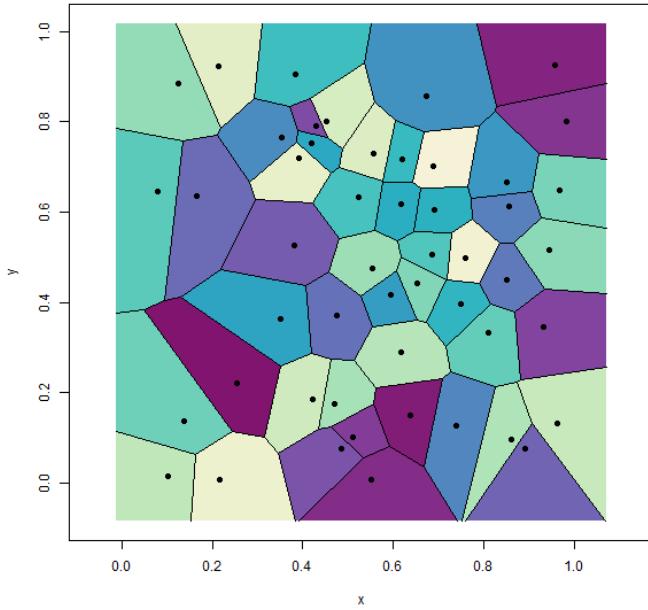


Figure 7.13: Thiessen polygons of 50 random points. Pickell, CC-BY-SA-4.0.

Thiessen polygons are a product of Delaunay triangulation described in the previous section. Figure 7.14 shows the relationship between the points, triangulation, circumcircles, and the Thiessen polygons. Connecting the circumcentres of the circumcircles produces the Voronoi diagram (Figure 7.14).

### 7.7.6 Centroids

A **centroid** is a point that represents the geometric centre of a polygon. For convex polygons, the centroid will always lay within the polygon, but for concave polygons, the centroid may lay outside the polygon (Figure 7.15). Circular polygons always have centroids that are equidistant to the boundary of the polygon (Figure 7.16).

### 7.7.7 Convex hull

A **convex hull** is the smallest polygon that contains some set of points. It is sometimes also referred to as a “convex envelope” or “convex closure” because

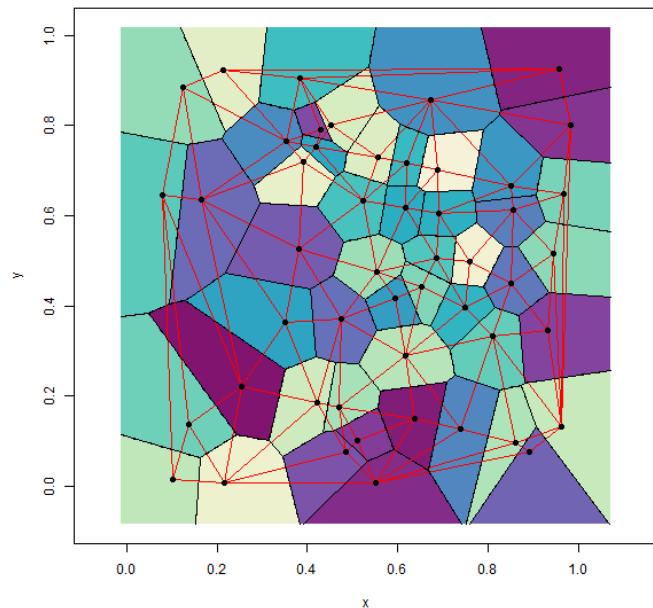


Figure 7.14: Delaunay triangulation (red lines) overlaid onto the Thiessen polygons. Pickell, CC-BY-SA-4.0.

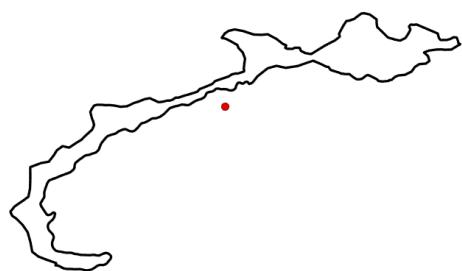


Figure 7.15: Concave polygon with the centroid (red dot) laying outside its boundary. Pickell, CC-BY-SA-4.0.

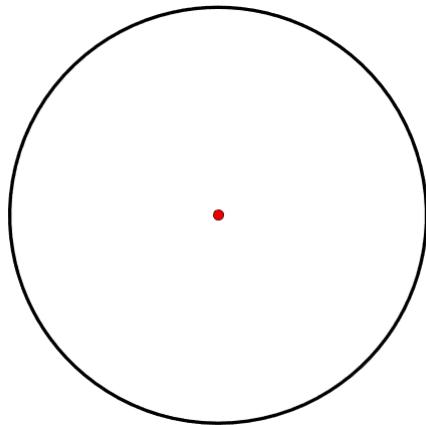


Figure 7.16: Circle polygon with the centroid (red dot) laying equidistant from the boundary of the polygon. Pickell, CC-BY-SA-4.0.

the perimeter of the polygon is formed by connecting the outermost points and closing or enveloping the remaining points. The convex hull is therefore the mathematical implementation of **topological closure**, where closure refers to the smallest closed set of points that contain the set of points. In practice, the convex hull is a bit like applying a rubber band around the outermost points so that the tension of the rubber band forms straight lines between the pairs of points in the closed set (Figure 7.17). There are several algorithms for computing the convex hull, including Jarvis' March [Jarvis, 1973], Graham's Scan [Graham, 1972], quickhull [Barber et al., 1996], and CudaHull [Stein et al., 2012].

Convex hulls are easily drawn by hand and are used for identifying a natural boundary for a sample set of points. Formally, the calculation is

### 7.7.8 Convex alpha hulls and alpha shapes

Convex hulls can be generalized to the concave case, called **convex alpha hulls** or **-shapes (alpha shapes)**, by adjusting the maximum radius of the circumcircles through a parameter, alpha  $\alpha$ . The objective of a convex alpha hull is to minimize the -shape formed by circumcircles of radius less than or equal to  $\alpha$ . Similar to the Delaunay triangulation, the circumcircles must be *open*, meaning they contain no other points in the set. Defined in this way, the final -shape may not result in closure of the full set of points and can result in holes where the distance between points exceeds  $2\alpha$ . Surprisingly, -shapes are prone to not existing at all. For the case of  $\alpha = 0$ , applying circumcircles of radius 0 results in an empty -shape and only the input set of points are returned without any boundaries. For  $\alpha = \infty$ , the -shape is equivalent to the

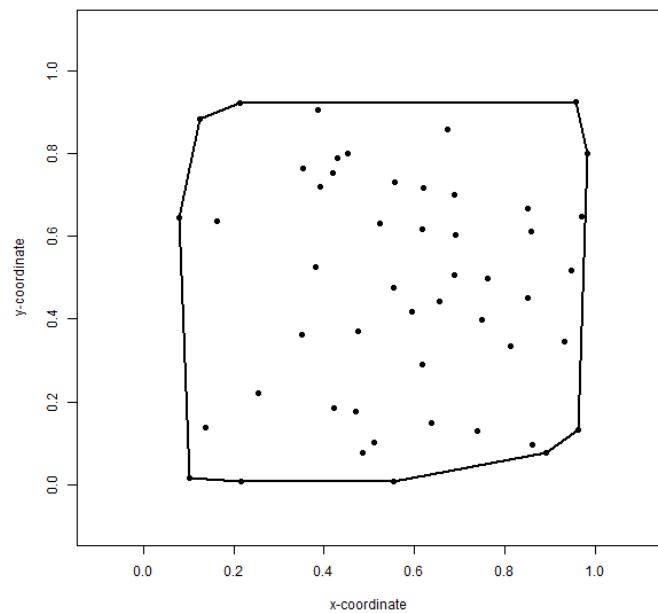


Figure 7.17: Convex hull formed by topological closure of the smallest closed set of points around the entire set of points. Arrangement of the points are the same as in the Thiessen polygons figure above. Pickell, CC-BY-SA-4.0.

convex hull because if we use circumcircles with an infinitely large radius, then all points in the set are bound to be enclosed by the resulting  $\alpha$ -shape, which like the convex hull must also minimize the bounding area. Figure 7.18 shows an animation of the  $\alpha$ -shapes for  $\alpha = 1$  to  $\alpha = 0$  for our set of 50 points.

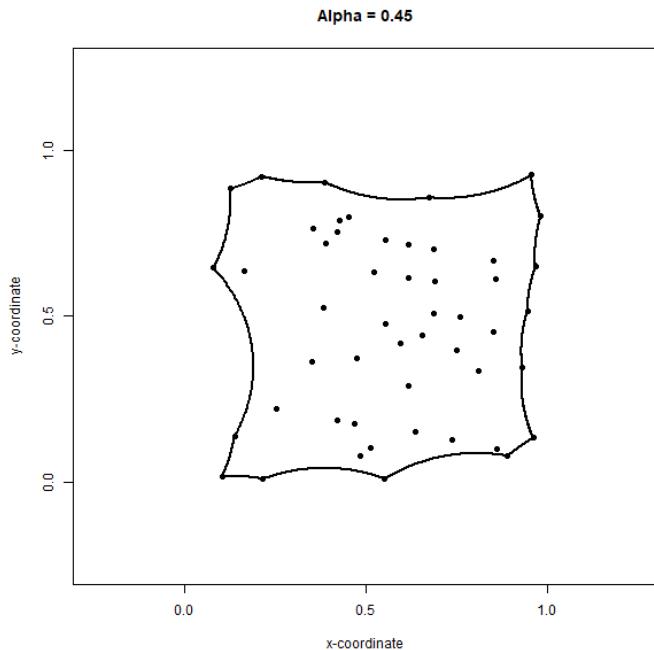


Figure 7.18: Concave alpha hull generates an alpha shape around a set of points. Online version of the figure is animated by alpha values from 0 to 1 by increments of 0.05. Pickell, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/topology-and-geocoding.html#fig:7-2d-alpha-hull>.

## 7.8 3D topologies

In this next section, a number of topologies are described that are important for 3D modeling and several examples are given using LiDAR (Light Detection and Ranging), which is the topic of Chapter 15. It is beyond the scope of this chapter to discuss the technology of LiDAR, so the reader is referred to Chapter 15 for a more in-depth discussion of LiDAR.

### 7.8.1 Multipatch geometries

Similar to multipart geometries, **multipatch geometries** associate several *faces* or *facets* to a single 3D feature such as a building or tree. In order for multipatch geometries to be topologically valid, they must form a closed set of faces, known as a **polyhedron**. Polyhedrons are comprised of *flat* faces that connect 3 or more vertices. Figure 7.19 illustrates the 5 Platonic polyhedrons, so-named after Plaot who initially wrote about them. The Platonic polyhedrons are a special type of *regular* polyhedron because they are the only polyhedrons that are highly symmetrical and have special transitive properties on the edges, faces, and vertices. As well, the Platonic polyhedrons are all examples of the 3-dimensional case of a convex hull (more on that in the next section). Most polyhedrons that we come across in environmental management like trees, lakes, glaciers, and buildings are very irregular and not Platonic.



Figure 7.19: The five Platonic solids are examples of regular, convex polyhedrons and multipatch geometries. From left to right: tetrahedron (4 faces); hexahedron (6 faces); octahedron (8 faces); dodecahedron (12 faces); and icosahedron (20 faces). Cyp and André [2005], CC-BY-SA-3.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomantics.ca/topology-and-geocoding.html#fig:7-platonic-polyhedrons>

### 7.8.2 3D Convex hull

The 3D convex hull is the smallest convex polyhedron that contains a set of 3D points (Figure ref(fig:7-3d-convex-hull)). The 3D problem is not unlike the 2D problem for finding the 2D convex hull, except instead of using 2D circumscribed circles we use 3D circumscribed spheres. Otherwise, the overall objective is the same, minimize the 3D polyherdon that encloses the 3D set of points. The 3D convex hull is frequently produced in order to generate a 3D object from a laser scan. Since points are 1-dimensional they have limited use beyond their enumeration within a volume or on a plane or on a line. By contrast, a 3D convex hull produces a polyhedron, which can be used to visualize the object that was initially scanned into a 3D point cloud. The 2D polygon faces that comprise a polyhedron can provide *shape* and interact with a simulated light source to improve perception of *depth*, which are qualities that are not provided by 3D points alone.

### 7.8.3 3D Convex alpha hull

The 3D convex alpha hull is Figure 7.20 below shows a 3D convex alpha hull for a deciduous tree near the

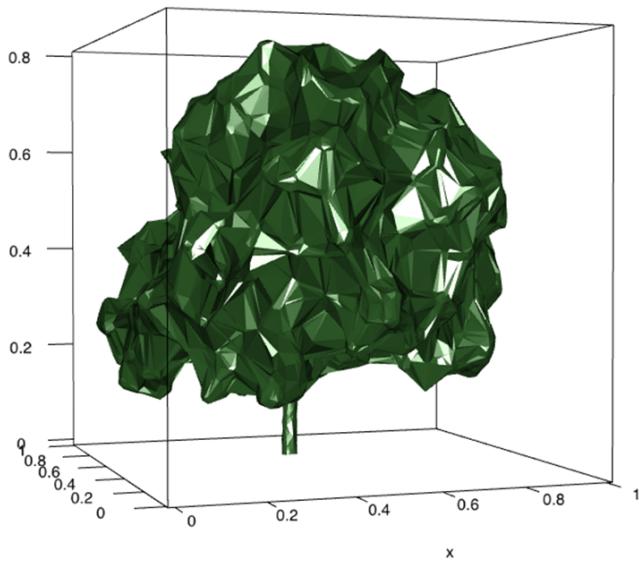


Figure 7.20: 3D concave alpha hull for a deciduous tree (species here). The alpha hull was generated using  $\alpha = 0.05$ . Data collected by Spencer Dakin Kuiper with a GeoSlam terrestrial laser scanner in Vancouver, Canada. Pickell, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/topology-and-geocoding.html#fig:7-3d-alpha-hull-deciduous>.

We can derive some useful information from these 3D convex alpha hulls. One obvious use of a polyhedron such as this is to calculate the volume, which for a tree might inform the amount of merchantable wood in the stem or the size of the canopy. Also, the polyhedron has replaced the points with polygon faces, which can be used to model shadowing and shading by simulating the position of the Sun in the sky.

### 7.8.4 3D Voronoi tessellation

Similar to Thiessen polygons, Voronoi tessellations can be undertaken for 3D point clouds. 3D Voronoi tessellations produce nearest neighbour polyhedrons around the space of each 3D point. Just as in the 2D case, these 3D Voronoi tessellations can be used to partition the 3D space. Figure 7.22 shows the use of a 3D Voronoi tessellation for use in mapping and visualizing polycrystals.

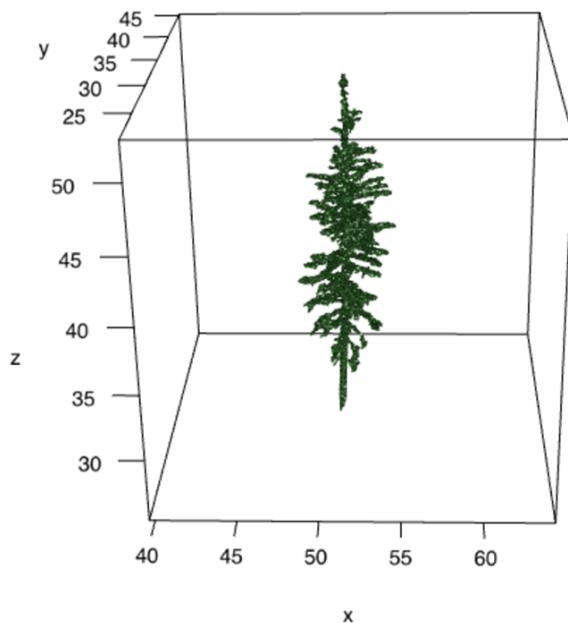


Figure 7.21: 3D convex alpha hull for a coniferous tree (*Abies lasiocarpa*). The alpha hull was generated using  $\alpha = 0.1$ . Data collected by Yangqian (Frederick) Qi with a terrestrial laser scanner in British Columbia, Canada. Pickell, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengis.ca/topology-and-geocoding.html#fig:7-3d-alpha-hull-conifer>.

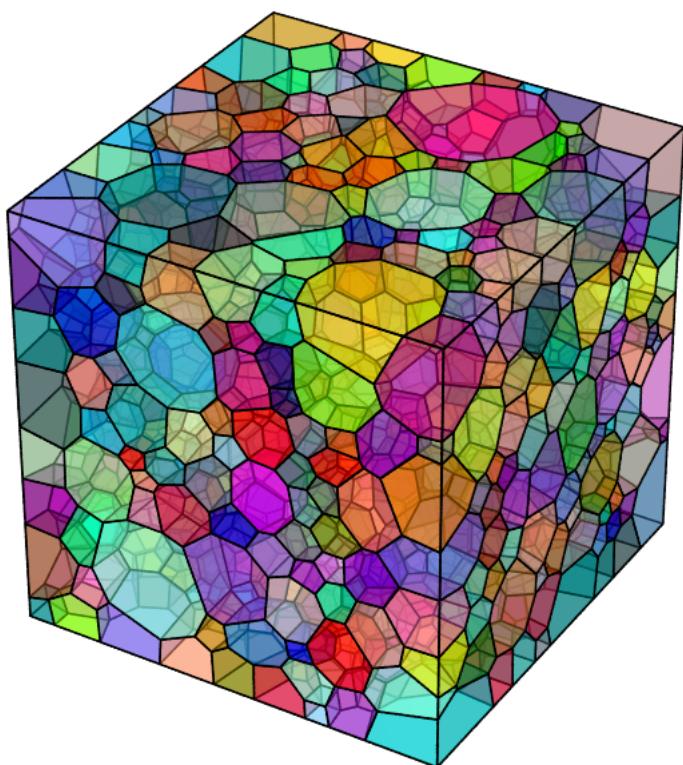


Figure 7.22: 3D Voronoi tessellation of simulated polycrystals. Quey, GPL 3.0.

## 7.9 Geocoding

**Geocoding** is the process of converting addresses to geographic coordinates, while **reverse geocoding** is the process of converting geographic coordinates to addresses (Figure 7.23). In order to achieve this conversion, an **address locator** uses reference spatial data that are mapped to a geographic or projected coordinate system in order to locate new addresses or coordinates.

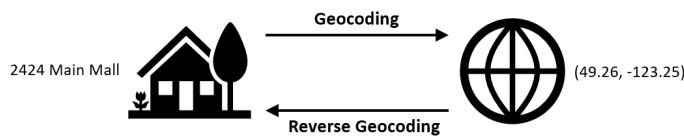


Figure 7.23: Conceptual figure showing the process of geocoding (converting addresses to geographic coordinates) and the process of reverse geocoding (converting geographic coordinates to addresses). Pickell, CC-BY-SA-4.0.

For example, consider that we are looking the 100-block of Main Street in Anytown, Canada. Neighbourhood blocks usually demarcate anywhere between 100 and 1000 unique civic numbers along a street segment. So the 100-block of our conceptual Main Street has addresses in the range of 100-199 (Figure 7.24). It is important to recognize that this is only a segment of Main Street, which presumably extends farther with additional segments for the 200-block, 300-block, and so on. (Remember, with proper planar topology, a single street can be comprised of many segments due to intersections with other streets.)

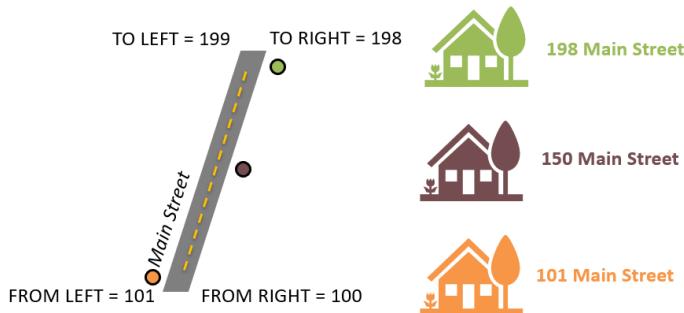


Figure 7.24: The 100-block of Main Street represents all civic addresses in the range of 100-199. Pickell, CC-BY-SA-4.0.

Suppose we have three addresses that we want to locate geographically along this segment: 101 Main Street; 150 Main Street; and 198 Main Street. If this segment of Main Street is mapped in a GIS, then we know the exact geographic coordinates (i.e., latitude and longitude) of the vertices and nodes (ends) of the street segment. Within the attribute table for this segment, we would find four

fields: FROM\_LEFT; TO\_LEFT; FROM\_RIGHT; and TO\_RIGHT (shown below).

STREET_NAME	FROM_LEFT	TO_LEFT	FROM_RIGHT	TO_RIGHT
Main Street	101	199	100	199

These fields indicate the range of civic numbers and the side of the street segment that a particular range falls on. The typical convention for address assignment within municipalities in Canada is odd-numbered civic numbers are on one side and even-numbered civic numbers are on the other side. In the GIS, these are arbitrarily assigned as *RIGHT* or *LEFT* sides, but geographically these addresses will occur on the North, East, South, or West “sides” of the street segment.

We can see that the values on the *LEFT* side of the street range from 101-199, which are odd-numbered, and the values on the *RIGHT* side of the street range from 100-198, which are even-numbered. Since the civic numbers of the street segment are known at the nodes (i.g., 100 and 101 at one end and 198 and 199 at the other end), then we can simply interpolate for any other civic number along the segment and identify the location of our three addresses (Figure 7.24). This interpolation process only places an address on the line segment (i.e., the centre of the street), so the locator must also geographically place the address on the correct side of the street using some offset value (usually in meters) that is usually perpendicular to the street segment.

### 7.9.1 Geocoding Assumptions and Limitations

One problem might seem obvious here: many cities have a street named Main Street. Therefore, an address locator relies on several pieces of reference spatial data such as maps of road networks, postal codes, cities, provinces or states, and countries. The address locator then works to *match* the input address against this database of spatial reference data. Thus, geocoding is both imprecise and inaccurate because the address locator relies on several assumptions. The primary assumption is that the input address exists and contains no typos or errors. Data entry by humans is a frequent source of typos and different styles for abbreviations (e.g., “E”, “E.” and “East”). An address locator can still geocode an address that does not exist as long as it is specified correctly, which results in an inaccurate location. If the input address is correct, but incomplete (e.g., “Main Street, Vancouver” is missing civic number, city, and province), then the address locator must match the other provided information against the spatial database (e.g., street name and city), which results in an imprecise location.

In addition to a set of geographic coordinates, one key result from geocoding an address is the **match score**, which is an indication of how well the address locator was able to match the address against the spatial reference database. The match score usually ranges from 0% (no match) to 100% (perfect match) and the calculation varies depending on how you want to penalize incomplete or

incorrect input addresses. Although it is frequently presented as a percentage, the match score is *not an indication of accuracy* and it really only reflects the confidence by the address locator given the reference spatial database. In other words, a completely inaccurate road network with correct names and civic numbers can conceivably “locate” an address with a 100% match score, but very low accuracy. The final limitation is that you cannot geocode addresses outside of the extent of the spatial data provided to the address locator. For geocoding over large areas, we often rely on geocoding services described in the next section.

### 7.9.2 Geocoding Services

If you are aiming to geocode addresses in a single city, then it is feasible to manually specify your own address locator using available spatial data such as roads, parcels, neighborhoods, and postal codes. However, for geocoding across large areas, this may not be feasible and you may instead rely on geocoding services that use large databases of reference spatial data. Commercially-available geocoding services are frequently used to provide routing, like Google Maps and Waze. However, these geocoding services do not provide match scores or any other indication of how confident or reliable the matches are.

## 7.10 Case Study: Working with Canadian Census Data

2016 Census Profile data

2016 Census Boundaries

## 7.11 Summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut in dolor nibh. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent et augue scelerisque, consectetur lorem eu, auctor lacus. Fusce metus leo, aliquet at velit eu, aliquam vehicula lacus. Donec libero mauris, pharetra sed tristique eu, gravida ac ex. Phasellus quis lectus lacus. Vivamus gravida eu nibh ac malesuada. Integer in libero pellentesque, tincidunt urna sed, feugiat risus. Sed at viverra magna. Sed sed neque sed purus malesuada auctor quis quis massa.

## Reflection Questions

1. Give some examples of situations where you might use planar and non-planar topology.
2. What are some examples of applications of geocoding and reverse geocoding?

3. Define topology.

## Practice Questions

1. Draw a convex hull for the points below.
2. Given ipsum, solve for lorem.
3. Draw ipsum lorem.

## Recommended Readings

Barber, C.B., Dobkin, D.P., and Huhdanpaa, H. 1996. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4): 469-483. <https://doi.org/10.1145/235815.235821>

Delaunay, B.N. 1934. Sur la Sph'ree Vide. *Izvestia Akademia Nauk SSSR*, VII Seria, Otdelenie Matematicheskii i Estestvennyka Nauk, 7:793–800.

Graham, R.L. 1972. An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set. *Information Processing Letters*, 1(4): 132–133. doi: 10.1016/0020-0190(72)90045-2.

Jarvis, R.A. 1974. On the identification of the convex hull of a finite set of points in the plane. *Information Processing Letters*, 2(1):18-21. [https://doi.org/10.1016/0020-0190\(73\)90020-3](https://doi.org/10.1016/0020-0190(73)90020-3)

Okabe, A., Boots, B., and Sugihara, K. 2007. Nearest neighbourhood operations with generalized Voronoi diagrams: a review. *International Journal of Geographical Information Systems*, 8(1): 43-71. <https://doi.org/10.1080/02693799408901986>

Stein, A., Geva, E., and El-Sana, J. 2012. CudaHull: Fast parallel 3D convex hull on the GPU. *Computers and Graphics*, 36(4): 265-271. <https://doi.org/10.1016/j.cag.2012.02.012>



# Chapter 8

# Network Analysis

Written by Nicholas Martino and Paul Pickell

Networks are **abstract structures** commonly used to represent patterns of relationships among sets of various *things* [Ajourlou, 2018]. Such structures can be used to represent social connections, spatial patterns, ecological relationships, etc. In GIS, the elements that compose geospatial networks are **geolocated** – in other words: they have latitude and longitude values attached to them. Network analysis encompasses a series of techniques used to interpret information from those networks. This chapter introduces basic concepts for building, analyzing and applying spatial networks to real-world problems.

## Learning Objectives

1. Understand what networks are and to identify the elements that compose them
2. Categorize different types of networks according to their topologies
3. Create spatial networks and learn how to apply them in various applications
4. Extract relevant information from spatial networks about the relationship between their elements, such as routes, distances and centralities

## Key Terms

Network analysis, Spatial networks, Graph theory

## 8.1 Introduction to Graph Theory

Graphs are the abstract language of networks [Systems Innovation, 2015a]. Graph theory is the area of mathematics that study graphs. By abstracting networks into graphs, one is able to measure different kinds of indicators that represents information about relationships that exist within a certain system. Why abstracting real-world elements into networks can be useful? Network analysis facilitates the study of data sets that demand information about their behaviour in terms of connectivity, flows, direction or paths. This is especially useful to understand the behaviour of complex adaptive systems such as societies, cities, ecosystems, etc. All graphs are composed of two parts: **nodes** and **edges** (or links).

## 8.2 Nodes

A **node** (or vertex) may represent any thing that can be *connected* with other things. For example, it can represent people in social networks, street intersections in road networks, or chemical compounds in molecular networks, among others.

## 8.3 Edges

**Edges** (or links), on the other hand, represent how vertices are interconnected to each other. So it may represent the vertices' social connections, street segments, molecular bindings, etc. The graph below represents rapid and frequent transit lines in Metro Vancouver. Each node represents a transit line and the edges represents connections between those lines.

## 8.4 Connectivity and Order

There are two major types of connections within the graphs: **directed** and **undirected**. Connections are directed when they have a specific node of origin and destination.

## 8.5 Direct

Directed graphs are networks where the order of elements change relationships between them. We represent directed connections with an arrow. The network below represents relationships between characters of Les Miserables. For example, in the case of the transit network we could use a directed graph to represent the path one has to take in order to shift from one line to another.

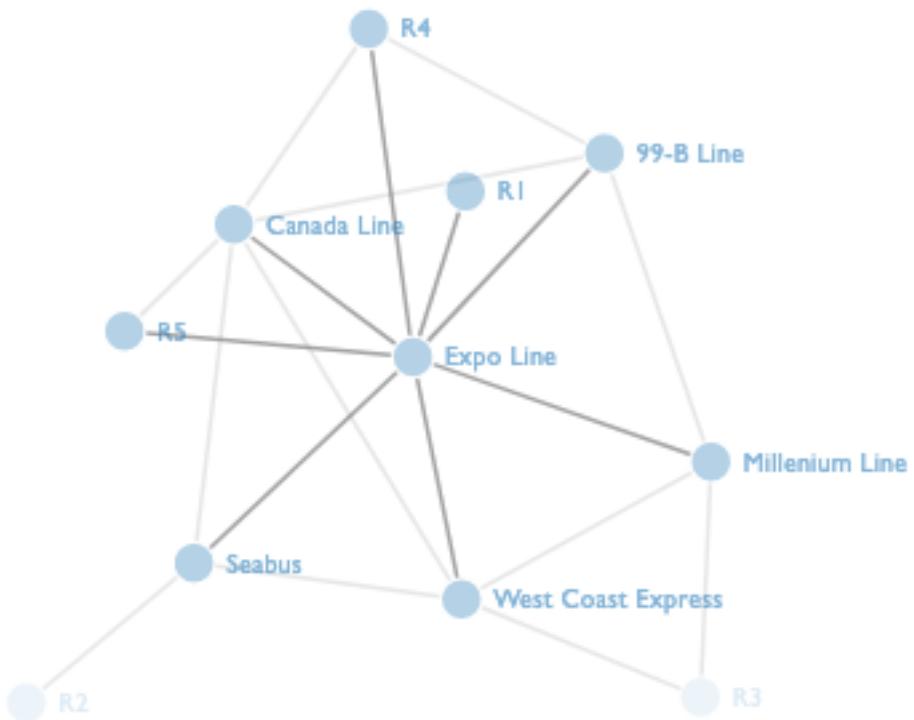


Figure 8.1: Graph representing connection between Metro Vancouver rapid and frequent transit lines. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengis.ca/network-analysis.html#fig:8-vancouver-transit-graph>.

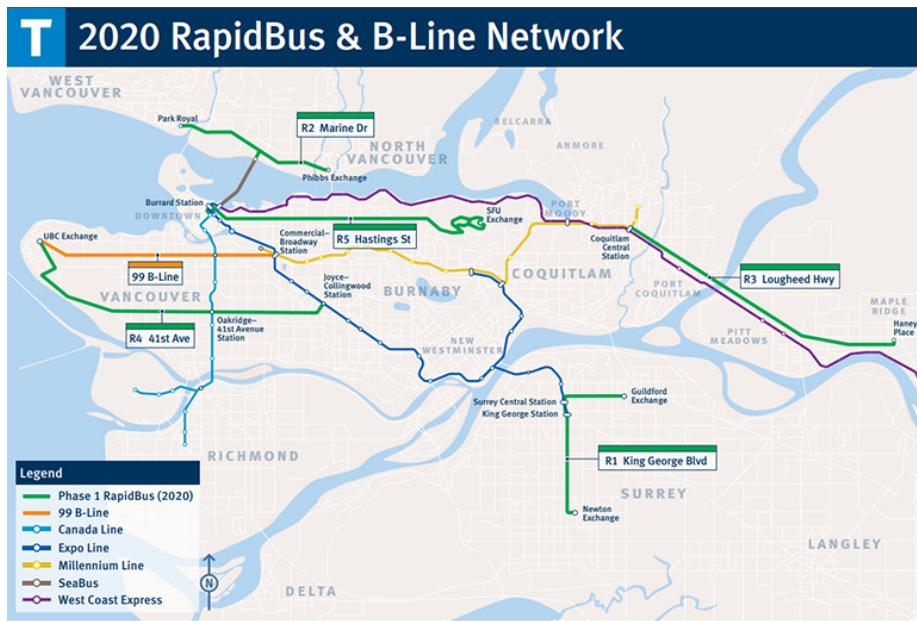


Figure 8.2: Rapid and frequent transit network in Metro Vancouver [TransLink, 2020].

## 8.6 Undirect

On the other hand, in an undirected graph, connections are represented as simple lines instead of arrows. The order of elements does not matter.

## 8.7 Network Topologies

Topology is the study of how network elements are arranged. The same elements arranged in different ways can change the network **structure** and **dynamics**. A very common example is the arrangement of computer networks.

## 8.8 Physical vs. Logical Topology

In GIS we use networks to represent spatial structures of various kinds. While all networks can be represented in an abstract space - this is, without a defined position in the real-world - some network analysis might be more useful when we attach physical properties to them, such as latitude and longitude coordinates. We call **logical topology** the study of how network elements are arranged in this abstract space. On the other hand, **physical topology** refers to the arrangement of networks in the physical space. We can then classify “types” of networks according to the way their nodes are arranged.

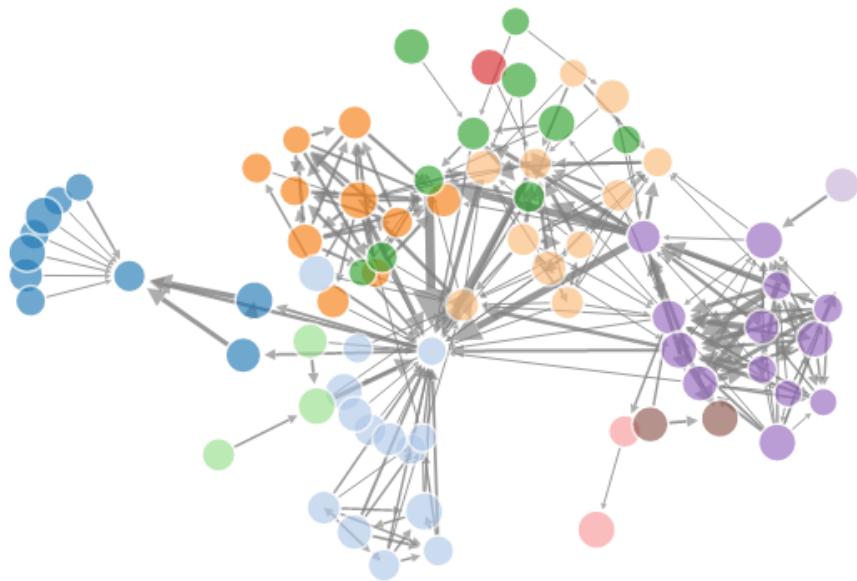


Figure 8.3: Example of directed graph for social relationships. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-directed-graph>.

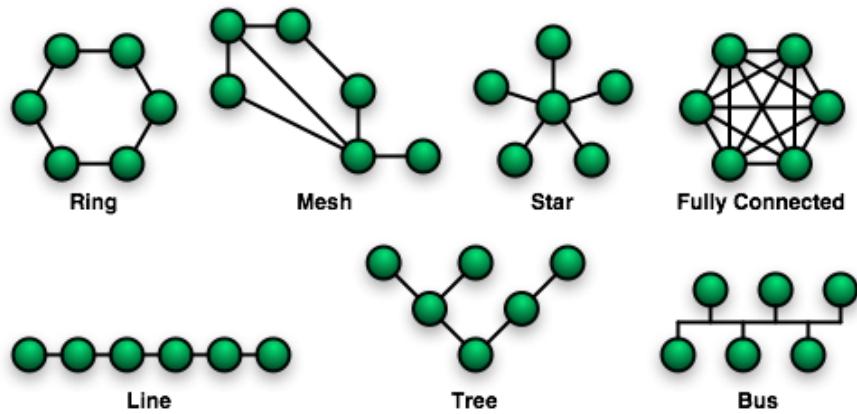


Figure 8.4: Abstract examples of network topologies [Wikibooks, 2018].

## 8.9 Non-Hierarchical Topologies

### 8.9.1 Lines

Lines are when nodes are arranged in series where every node has *no more than two connections*, except for the two end nodes. A rail transit line, for example, can be represented as a line network. The map below portrays the SkyTrain Millennium Line in Vancouver. Each node represents a stop and the lines the connections between those stops.

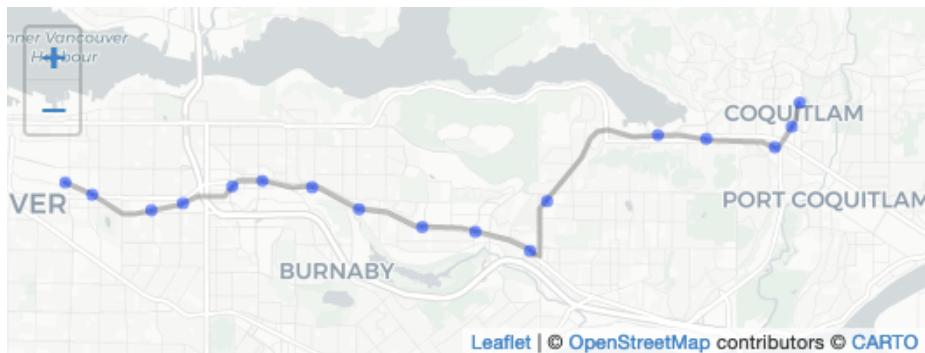


Figure 8.5: Rail transit line in Vancouver. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: [https://www.opengo-matics.ca/network-analysis.html#fig:8-vancouver-rail-transit](https://www.opengeo-matics.ca/network-analysis.html#fig:8-vancouver-rail-transit).

### 8.9.2 Rings

Rings are similar to lines except that there are no end nodes. So each and every node has *two connections and the “first” and “last” nodes are connected to each other* forming a circle. The spatial structure of the Stanley park seawall trail in Vancouver resembles a ring. In this example, nodes stand for intersections and view spots and edges are the connections between these spots along the seawall.

### 8.9.3 Meshes

In a mesh, *every node is also connected to more than one node*. However, in this case nodes can be connected to more than two nodes. Connections in a mesh are non-hierarchical. Contrary to rings and lines where there is only one possible route from one node to another, in a mesh there are multiple routes to access other nodes in the network. A common way to generate a mesh network is using **Delaunay triangulation**, where nodes are connected in order to form triangles and maximize the minimum angle of all triangles [Wikimedia, 2021a]. Mesh configurations are commonly used in decentralized structures such as the internet.

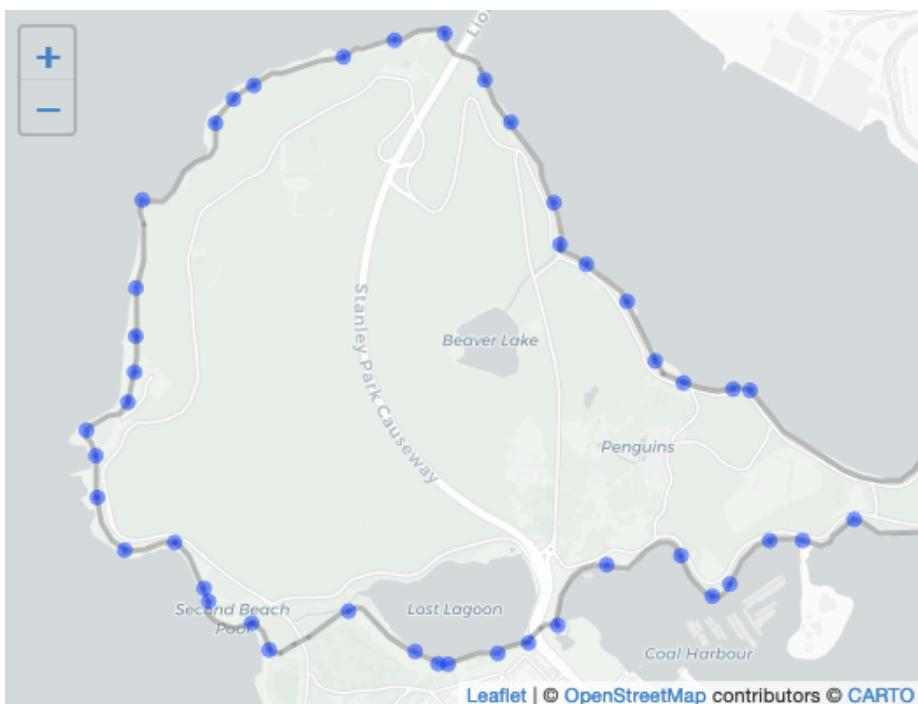


Figure 8.6: Ring of the Stanley Park seawall. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-stanley-park-seawall>.

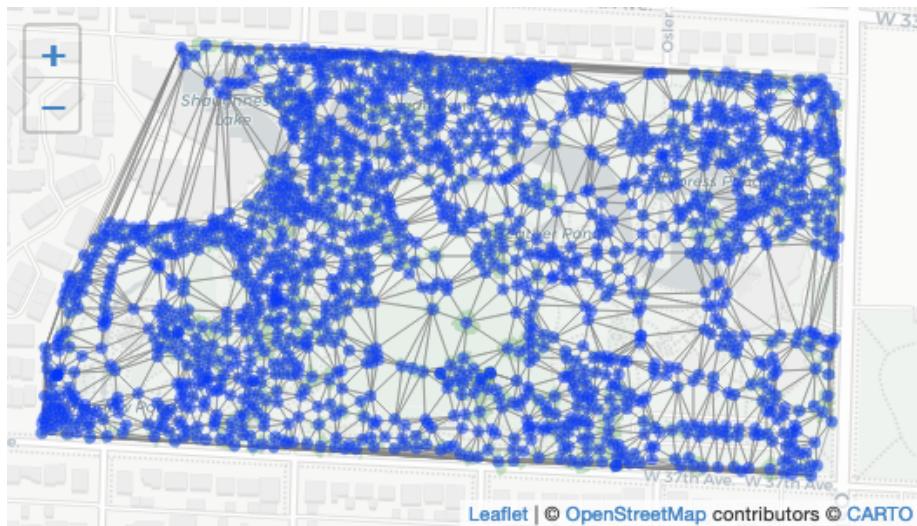


Figure 8.7: Tree canopy mesh. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-tree-canopy-mesh>.

#### 8.9.4 Fully Connected

As the name suggests, in fully connected networks *every node is connected to every other node*. The graph representing all possible origin-destination commutes among Metro Vancouver municipalities is a type of fully connected network.

### 8.10 Hierarchical Topologies

Different from non-hierarchical topologies, hierarchical configurations are structured around a central node or link. By looking into hierarchical topologies it becomes easier to understand the notion of depth. The more distant a node is from the central node or link, the more depth it has. Hover the mouse over the nodes in the following maps to check out their depth.

#### 8.10.1 Stars

Stars are hierarchical structures where *two or more nodes are directly connected to a central node*. This concentric garden at the University of British Columbia can be represented according to a star topology.

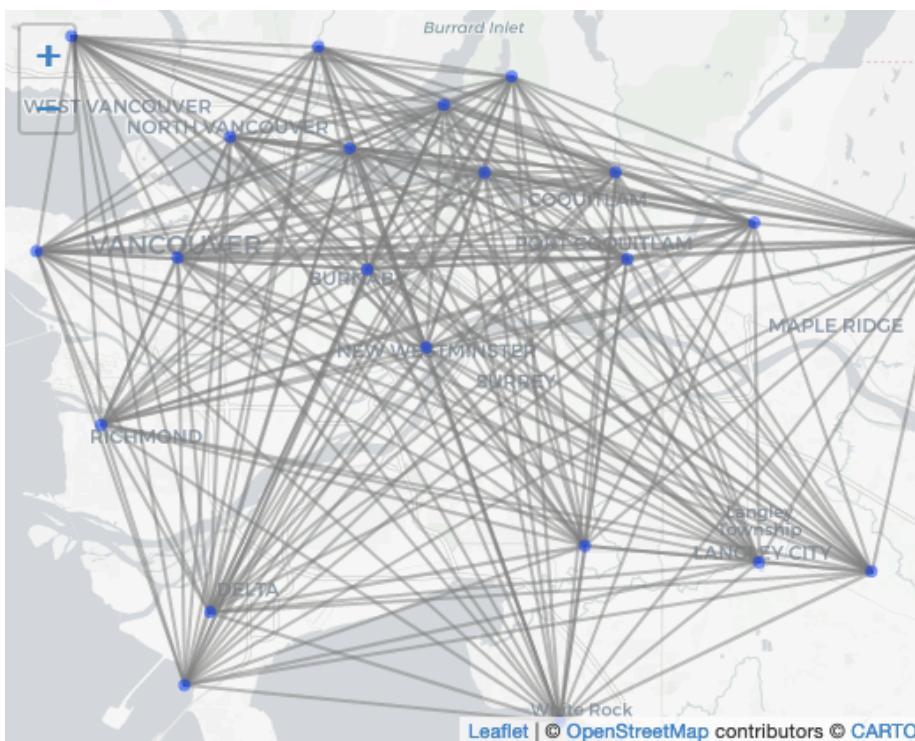


Figure 8.8: Possible origin-destination commutes between municipalities within Metro Vancouver. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-tree-canopy-mesh>.

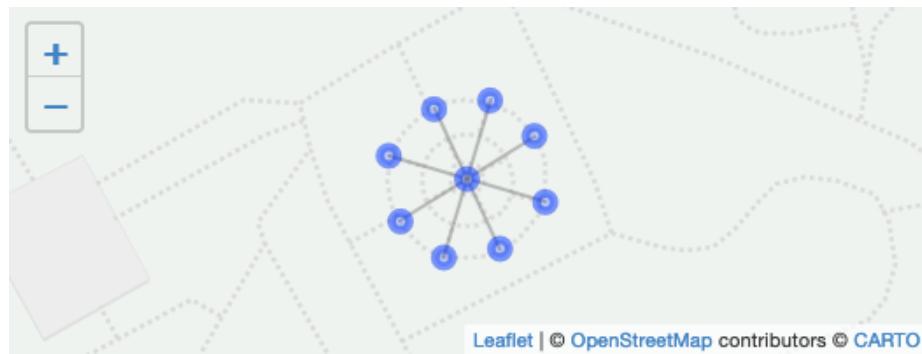


Figure 8.9: Spatial structure of a concentric garden at UBC. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-spatial-structure-garden>.

### 8.10.2 Buses

Buses are structures where *every path from one node to another passes through a central path or corridor*. If we isolate a street segment from an urban street network, the connections between buildings and streets depict a bus topology.

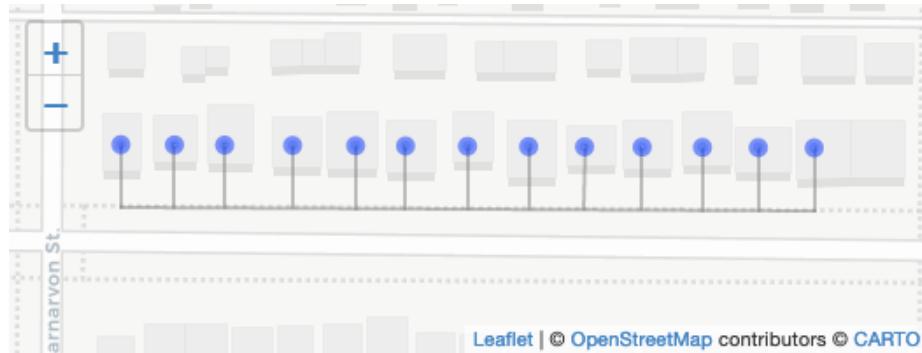


Figure 8.10: Connections between houses and streets. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-connections-houses-streets>.

### 8.10.3 Trees

In tree topologies, nodes are *structured from a root node and arranged into edges* that are similar to branches of a tree. This highly hierarchical structure creates

a sort of **parent - child** relationship amongst nodes. The spatial configuration of boat marinas are usually structures in tree-like topologies. By definition, all tree network structures will always have more than one terminal nodes (a node that only has one connection to the network).



Figure 8.11: Tree spatial structure of a boat marina. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-boat-marina>.

## 8.11 Spatial Network Analysis

Networks can then be arranged according to various different configurations. Aside from classifying networks into different types according to their topologies, some of the most useful features of spatial network analysis refers to how to **extract information** from these structures given certain parameters.

## 8.12 Network Tracing

The act of modelling spatial networks is called **network tracing**. When tracing a network it is important to bear in mind the **direction** with which information is added to the network, especially when this orientation information is important to further analyze **flows** and relationships within such structure. For example, when mapping hydrological networks to study its flows it might be useful to model the direction of streams coherently as this might be an important information to represent the dynamics of the network.

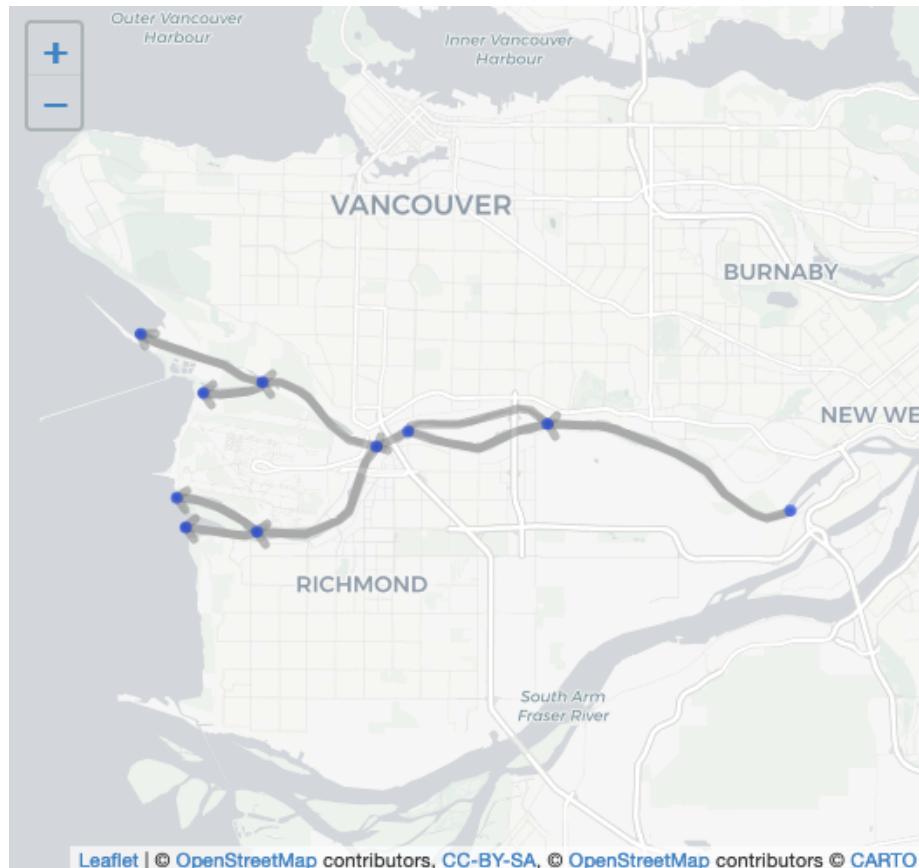


Figure 8.12: Graph representing Fraser River Flows. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-fraser-river-flows>.

## 8.13 Linear Referencing

Linear referencing is a method of using geographic locations for measuring relative positions along a linear feature. In network analysis, linear referencing techniques can be used for finding the length of paths along the network [Ramsey, 2012]. In this method, the graph elements are defined in terms of their physical location and edges are used to calculate distances among parts of the network.

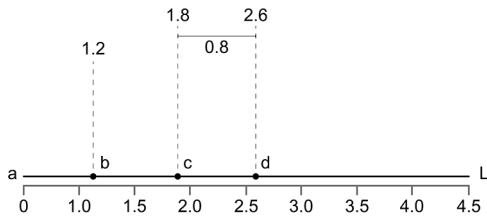


Figure 8.13: Example of using linear referencing to measure distance between points.

In the above figure we can see how linear referencing systems work. Considering one would like to measure distances from node  $a$  along a network link  $L$ , distance measures can be used to locate, for example, points  $b$ ,  $c$  and  $d$  along line  $L$ .

## 8.14 Routing

One application of linear referencing is to find routes between nodes is an useful application of spatial networks. This is how mapping tools help us navigate the world by finding the most efficient route to move around the city, for example. [Systems Innovation, 2015b].

## 8.15 Least Cost Paths

Usually multiple paths can be traced along a network to go from one point to another. The notion of **cost** allow us compare the degree of *difficulty* needed to cross such paths. With this information, it is possible to rank different routes. In spatial networks, cost usually relate to the necessary distance (either physically or logically) to go from a certain node to another, but they might also represent other aspects such as time, traffic, elevation, current flows, etc. For example, way finding tools that are commonly used to help us to locate and move around in the city usually takes into account multiple costs such as distance, traffic and/or elevation. The **least cost path** is the *easiest* way to go from one point of the network to another.

It can be found by associating the costs with elements of the network (either nodes or edges) and summarizing the total cost of certain routes. Usually **linear**

**referencing** techniques are used to calculate costs by storing locations along measured linear features. The map below displays the least cost path in terms of physical distance (shortest path) between two points.

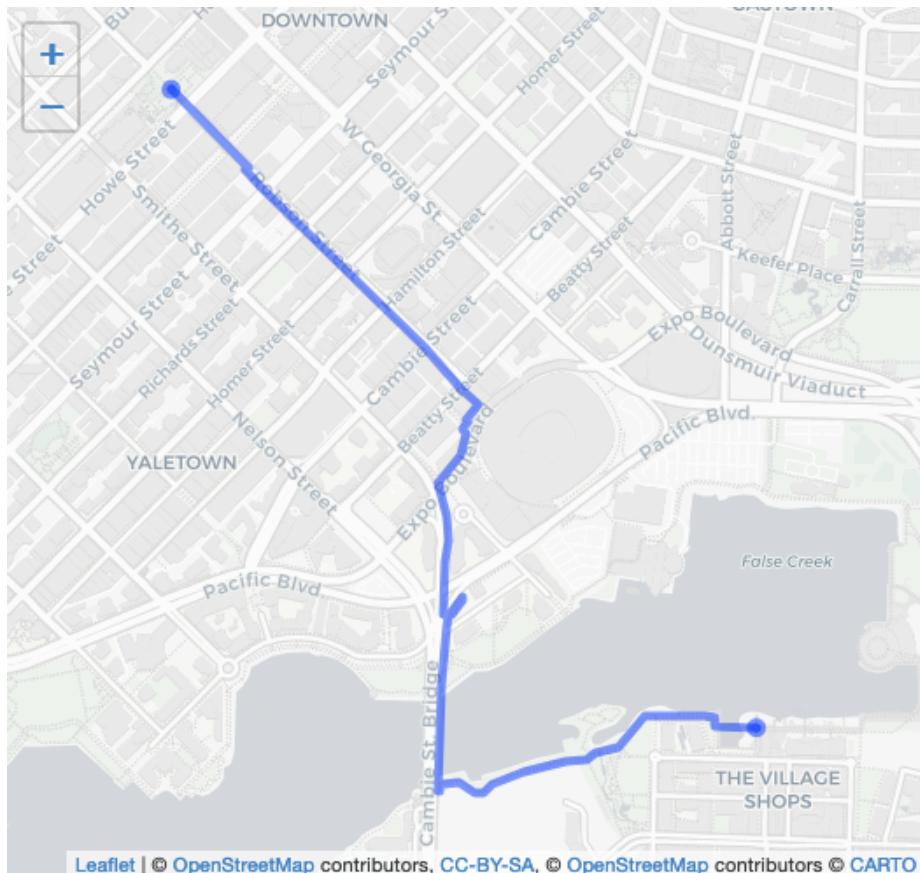


Figure 8.14: Least cost path (in terms of distance) between two points. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-least-cost-path>.

## 8.16 Least Cost Corridors

Although most network analysis techniques are suitable for vector data, raster layers can also be analyzed. One application of using network analysis with **raster data** is the finding of least cost corridors. While least cost paths helps to find linear paths along the network between two points, least cost corridors are based on the overlay of two *cost accumulative* rasters.

## 8.17 Reach Analysis

Reach techniques are commonly used to find the incidence of defined elements *within a certain radius from a chosen node*. All possible routes are modeled. The number of **terminal nodes** varies according to the network structure. Urban walkability indices usually uses reach techniques to assess the intensity of certain indicators (such as intersection density or non-residential land uses) given a walkable radius [Martino, 2020]. In the map below we portray the network reach from a given origin point into the Pacific Spirit Regional Park within 400m (red), 800m (yellow), 1200m (green) and 1600m (blue) radii.

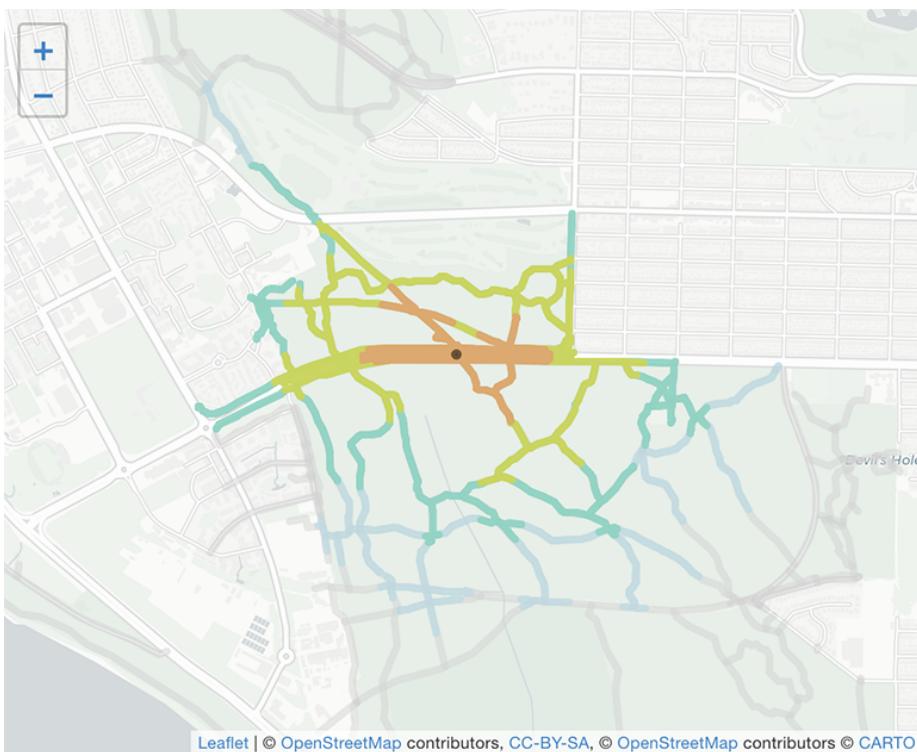


Figure 8.15: Reachable segments within multiple distance radii. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-reach-analysis>.

## 8.18 Network Centrality

Nodes and edges of a graph can also be ranked in terms of how **important** they are to the overall network. Network centrality measures represent whether elements of a graph are more central or peripheral to the overall system. Such

measures can then be interpreted as indicators of importance. Applications are endless. Centrality measures are used for ranking search engine pages [Wikimedia, 2021b], for finding persons of interest in social networks [Ajourlou, 2018] and for modelling movement in street network [Hillier et al., 1993]. There are several **centrality measures** that serve to the most various purposes. Some of the most commonly used ones are Closeness and Betweenness centrality.

## 8.19 Closeness Centrality

Closeness centrality measures *how close each node is to every other node of the graph* in terms of topological distances. It highlights nodes located in easily accessible spaces. For example when analyzing the closeness of street intersections at the University of British Columbia (UBC), intersections in core streets such as the Main Mall, Agronomy Road and Northwest Marine Drive are ranked as highly central whereas more residential and segregated areas such as Acadia Road are ranked with lower closeness centrality.

Closeness is calculated based on the **logical distance** from one vertex to all the other vertices in the network. The formula for estimating closeness centrality of a vertex  $i$  is:

$$c_i = \sum_j \frac{1}{d_{ij}}$$

where  $d_{ij}$  means the logical distance from  $i$  to  $j$ .

## 8.20 Betweenness Centrality

Betweenness centrality measures *how likely a node or an edge is to be passed through* when going from every node to every other node of the graph. If we imagine agents travelling from each node to every other node and back, betweenness centrality would be the trail left by those agents. While closeness highlights central spaces, betweenness highlights pathways that lead to those central spaces. Using the same street network at UBC we can calculate the betweenness of segments.

Betweenness is calculated based on the number of **shortest paths** (in logical distances) from all nodes to all other nodes. According to the documentation of the graph-tool software, betweenness of a vertex  $C_B(v)$  is defined as:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where  $(v)$   $\sigma_{st}$  represents the number of shortest paths from node  $s$  to node  $t$  and  $\sigma_{st}(v)$  represents the number of those paths that pass through  $v$  [graph-tool]. We can use centrality measures to evaluate how accessible certain spaces are from the point of view of their spatial structure with a broader system.

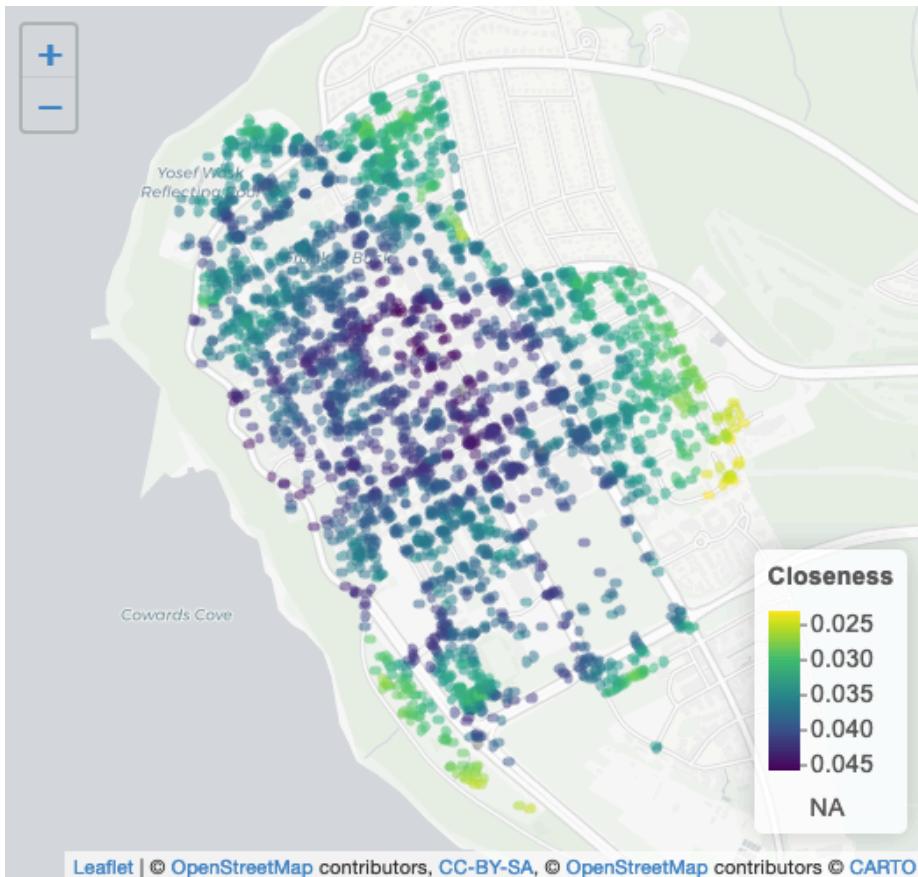


Figure 8.16: Closeness centrality of the street intersections at UBC. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-closeness-centrality-UBC>.

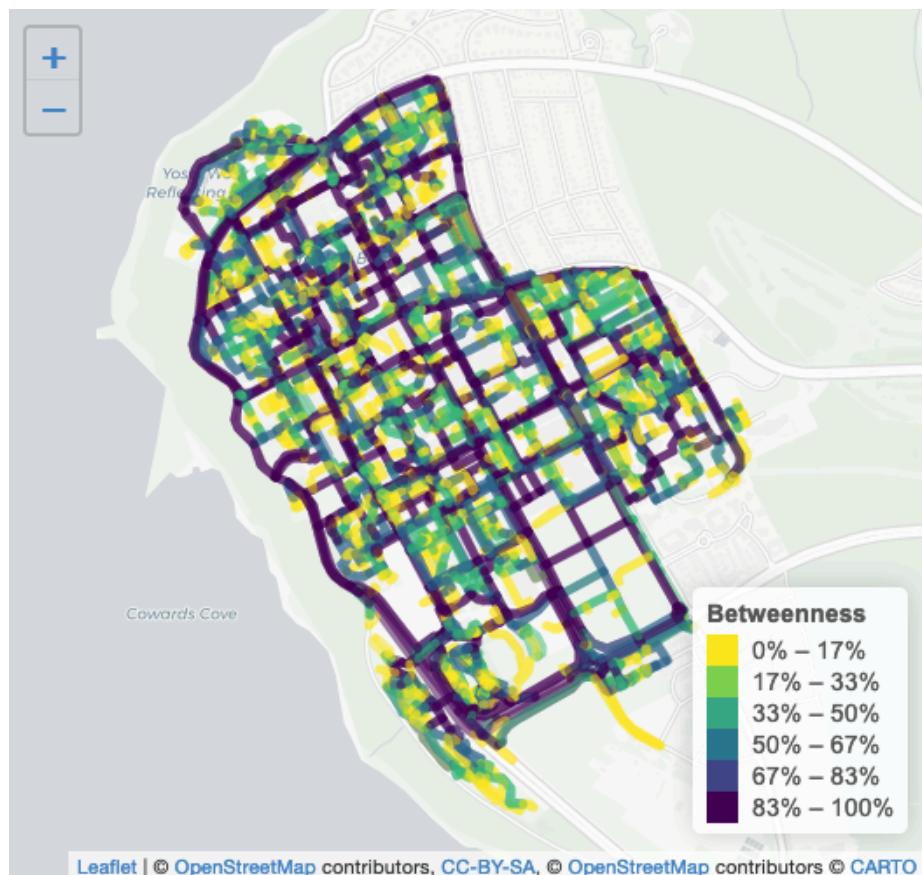


Figure 8.17: Betweenness centrality of street segments at UBC. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-betweenness-centrality>.

## 8.21 Case Study: Central and Peripheral Green Spaces in Vancouver

Are green spaces evenly accessible throughout the whole city? Which parks are topologically *closer* to the city as a whole? Centrality analysis of the street network can be used to answer these questions.

First we need to find the Closeness measure for the street network of the City of Vancouver. The network information was downloaded from OpenStreetMap. The software graph-tool was used to calculate the centrality measure. With the results of centrality for all street intersections in the city, we can overlay Parks & Green spaces data from the City of Vancouver Open Data portal and get the average closeness of nodes within each green space.

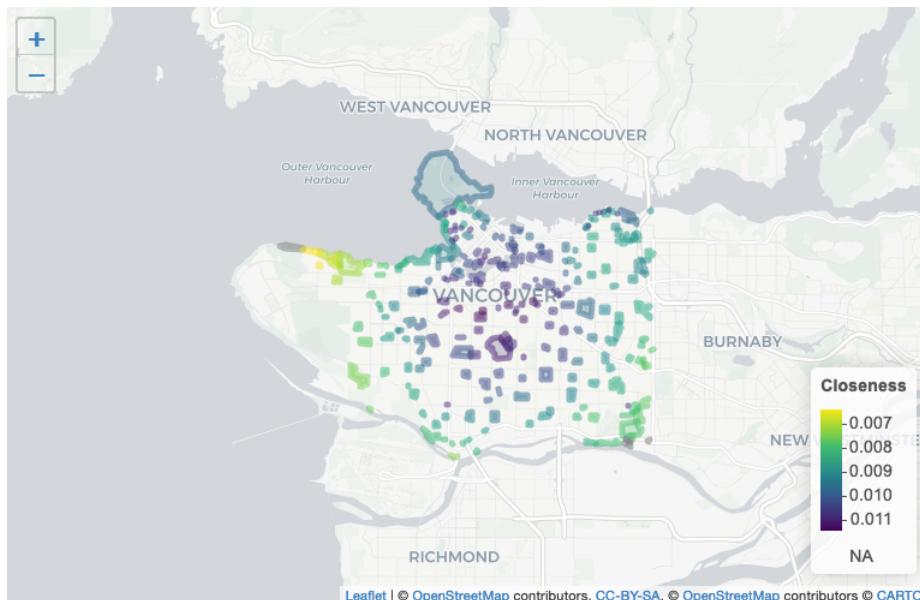


Figure 8.18: Closeness centrality of parks and green spaces at the City of Vancouver. Data from [City of Vancouver, b] and licensed under the Open Government License - Vancouver. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/network-analysis.html#fig:8-closeness-centrality-vancouver>.

Results show parks located in the middle of the city have **higher closeness** than parks located at the edges. In other words, these parks are located in parts of the city that have easy access to the city's street network as a whole. As the histogram leans towards the right, we can conclude that there are more parks with higher closeness than parks with lower closeness.

## Practice & Reflection

Some questions to reflect and better understand the basic concepts and applications of network analysis explained in this chapter:

- Which types of behaviour can be modelled and understood using network analysis techniques?
- What is the difference between *physical* and *logical* distances?
- How different *costs* can be used for routing along spatial networks?
- How can network centrality measures be interpreted in spatial networks?

# Chapter 9

# Raster Analysis and Terrain Modelling

Written by Paul Pickell

Many data are frequently collected and represented in a raster format. In this chapter, we will look at how rasters can be analyzed with a specific focus on deriving terrain information from elevation rasters. We will explore some useful properties of the raster structure that can be exploited for insightful analysis of land and water alike. Image-based raster analysis will be a later topic of Chapter 14.

## Learning Objectives

1. Understand the principles of raster analysis
2. Recognize the types of digital vertical models and their uses
3. Synthesize the terrain and hydrological information that can be derived from a digital elevation model
4. Apply raster functions in terrain and hydrology workflows
5. Visualize terrain with 3D geovisualization methods

## Key Terms

Digital Elevation Model, Digital Vertical Model, Digital Terrain Model, Digital Surface Model, Anaglyph, High-Pass Filter, Low-Pass Filter, Focal Function, Local Function, Global Function, Zonal Function, Derived Raster, Classified Raster

## 9.1 Raster Analysis

Raster analysis is the process of deriving, classifying, and combining raster data layers together. Although any given raster analysis may involve one or all of these steps, one thing that all raster analyses have in common is what is known as base data. **Base data** can be raster or vector data and represent some existing information about the area of interest. For example, you may have a digital air photo or other base map that you want to derive, classify, and/or combine with other information.

We can perform calculations on rasters to **derive** new rasters. For example, we could add, subtract, divide, or multiply a constant value to all cells in a raster. As well, we can calculate more complex indices or functions from a single raster, which we will see some examples of later in this chapter.

We can also **classify** rasters, a process that involves modifying the values of the cells in the input raster based on some conditions to produce a new raster with new values. Sometimes this process is also referred to as **reclassification**. For example, maybe we desire a raster with binary values of 1 or 0 based on the condition that the value of a raster cell is greater than 10 ( $=1$ ) or less than or equal to 10 ( $=0$ ). Then, if a cell value in the input raster is 9, then the value for that cell in the output raster will be 0, and so on.

Finally, after deriving and classifying our rasters, we often need to **combine** or overlay them in order to solve some problem. Suppose we have a binary raster that represents land ( $=1$ ) or water ( $=0$ ) and we have another binary raster that represents good planting conditions ( $=1$ ) and poor planting conditions ( $=0$ ). Then, if we multiply these two rasters together, we will identify land that has good planting conditions ( $=1$ ) across our area of interest. As you can see, combining raster data together can be a powerful tool for solving environmental management problems. Although it is common to reclassify rasters into binary rasters for a simple multiplication overlay like in the previous example, it is also possible to weight several rasters during combination and produce a continuous value in the output. We will see an example of this kind of combination in the case study.

## 9.2 Digital Vertical Models

**Digital Vertical Models (DVM)** represent vertical heights and elevations of terrain and features. As we learned from Chapter 2, *vertical* refers to the axis of geographic space that is orthogonal (i.e., perpendicular) to a vertical datum, such as a geoid or ellipsoid. Thus, a DVM may represent elevation of terrain or the height of a building. Both of these examples have specific meanings and calculations relative to the vertical datum, which we will explore in more detail in the following sections.

### 9.3 Digital Elevation Models (DEM)

As the name suggests, a **Digital Elevation Model (DEM)** is a digital representation of elevation data or heights above a vertical datum. We call it a *model* because, as you know from reading Chapter 3, all spatial data formats are inherently spatial data models. As is the case for all rasters, DEMs represent sampled data. That is, the elevations that are represented continuously in the raster DEM are in fact samples of elevation at the raster cell centres. For this reason, a DEM is a model that is simplifying elevation to some degree depending on the cell size.

Figure 9.1 shows a DEM for Mount Assiniboine near the Alberta and British Columbia border. Mount Assiniboine has a distinctive pyramidal peak that is not immediately apparent from the bird's-eye view of a DEM. It is common to see DEMs displayed in this black and white colour scheme, where black represents lower elevations and white represents higher elevations. Can you tell where the peak is? We will look at more sophisticated colour schemes and geovisualizations of terrain in a later section.

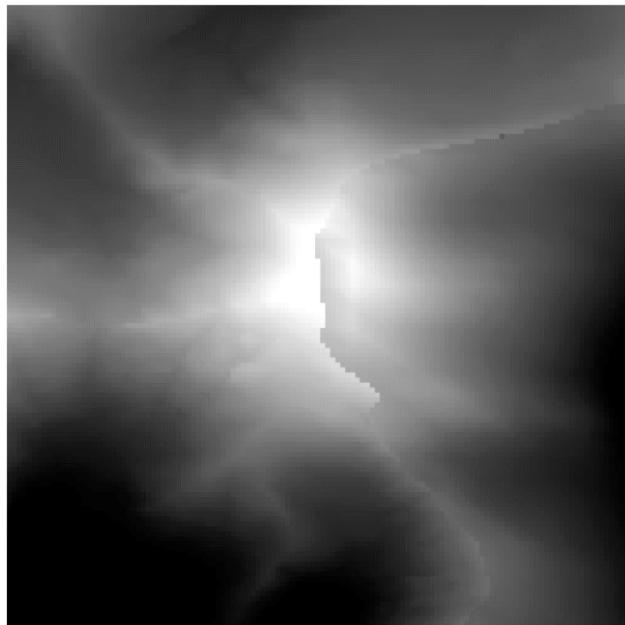


Figure 9.1: Digital Elevation Model of Mount Assiniboine at the border of Alberta and British Columbia, Canada. Data from Natural Resources Canada [2015]. Pickell, CC-BY-4.0.

Figure 9.2 shows another DEM, but this time for some the flattest terrain in

Canada over 1,600 km away from Mount Assiniboine on the Canadian Shield at the mouth of the Nelson River in Manitoba. The Nelson River drains an area of more than 1 million  $\text{km}^2$  of land across much of the Canadian prairie provinces and pours into Hudson Bay. The Canadian Shield is home to the oldest terrain on Earth and millions of years of erosion have reduced it to rolling hills and flattened horizons.



Figure 9.2: Digital Elevation Model of Nelson River pouring into Hudson Bay, Manitoba, Canada. Data from Earth Resources Observation and Science Center [2018]. Pickell, CC-BY-4.0.

Over an entire extent of  $3,300 \text{ km}^2$ , the Nelson River DEM in Figure 9.2 only varies by 44 m. Compare that with the 1,328 m of difference from the white peak of Mount Assiniboine to the black valleys of the Rocky Mountains over about  $5 \text{ km}^2$  in Figure 9.1. Throughout the remainder of this chapter, we will look at raster analysis applied to these two extreme examples of terrain.

## 9.4 Digital Terrain Models (DTM)

A **Digital Terrain Model (DTM)** represents elevation through points and lines and is often erroneously confused with a DEM. The vector-based data format allows elevation to be sampled at a higher density in areas where elevation changes quickly in space (e.g., Mount Assiniboine) and at lower density in areas

where elevation changes gradually (e.g., Nelson River). Lines can be used to model mountain ridges, river banks, fault lines, and coast lines where elevation might be constant and it would be useful to represent the elevation as a line feature instead of a point of elevation. It is important to recognize that DTMs can be converted to DEMs through a process of interpolation (more on that in Chapter 10), but a DEM cannot be converted to a DTM because a DEM is a regular grid or equally-spaced elevation samples.

## 9.5 Digital Surface Models (DSM)

Up to this point, we have been looking digital vertical models of *bare Earth*, that is, just plain old elevation of terrain above a vertical datum. If you want to represent the height of features above the bare Earth, like houses and trees, then you would need to use a **Digital Surface Model (DSM)**. Whereas a DTM and DEM both represent elevation above a vertical datum, which is usually mean sea level represented by the geoid, a DSM represents height above a DTM or DEM. In this way, we can model features on the surface and their heights. We will look at more examples of working with DSMs in Chapter 15.

## 9.6 Raster Functions

Raster functions are algorithms that perform operations on one or more cells of the raster to produce new calculations or derivatives. In the following sections we will explore four commonly used raster functions: local, focal, global, and zonal.

## 9.7 Local

A **local function** is the simplest to understand because each cell in the input raster is operated on independently of all other cells. Simple arithmetic operations such as addition, subtraction, division, and multiplication are all examples of focal functions when applied to a raster.

**Input Raster**

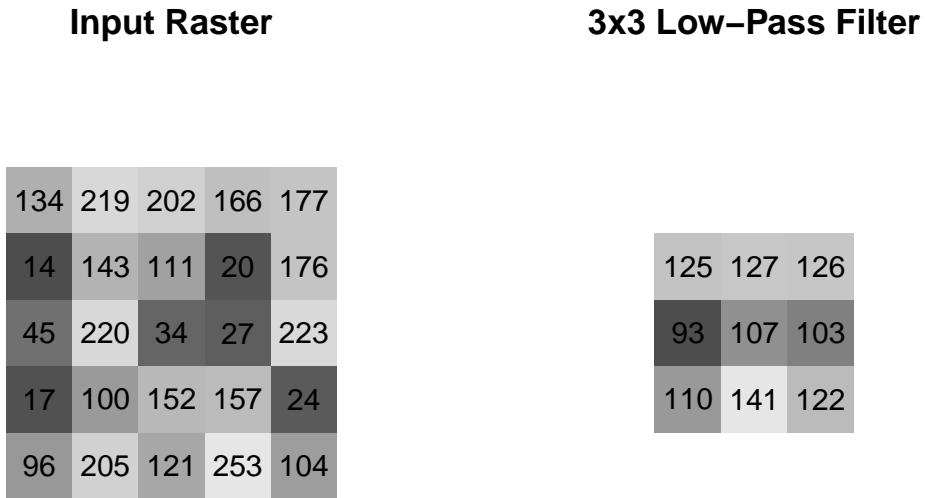
134	219	202	166	177
14	143	111	20	176
45	220	34	27	223
17	100	152	157	24
96	205	121	253	104

**Input Raster + 5**

139	224	207	171	182
19	148	116	25	181
50	225	39	32	228
22	105	157	162	29
101	210	126	258	109

## 9.8 Focal

A **focal function** takes full advantage of the raster data model by moving a window or kernel over the raster in order to calculate new values. The kernel must always be a square with odd-numbered dimensions such as 3x3 or 11x11. This ensures that there is a focal cell at the centre of the kernel that determines which cell in the output will take on the value of the operation. In practice, the kernel is moved column-by-column, row-by-row over the entire raster and a calculation is performed using the values of the input raster that coincide with the kernel. The kernel itself is also comprised of weights that, when multiplied against the input raster at a given location, yields a set of values that can be summed or averaged. Thus, the operation of the focal function can take many different forms such as calculating the mean, minimum, maximum or any other operation over the kernel. As a simple example, a mean focal function with a 3x3 kernel will calculate the mean value of all 3x3 cell neighborhoods in the raster yielding the result below.



Mathematically, the weights of the 3x3 kernel are all 1's so that when it is multiplied against a particular location on the input raster the values of the input raster are returned.

$$k = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

The mean is then calculated by adding all the values together and dividing by the number of cells in the kernel. You will see that our input raster has shrunk down from dimensions of 5x5 to 3x3 after applying the 3x3 kernel. This is due to the fact that there are insufficient cells along the edge and corners of the raster to divide by the number of cells in the kernel. One solution is to use padding, which simply adds 0's along the outside of the input raster. Padding is not always an elegant solution because you will still have edge effects, but you will at least maintain the dimensions of your input raster in the output. Below is the result of the same focal function, but with padding of 0's.

**Input Raster**

134	219	202	166	177
14	143	111	20	176
45	220	34	27	223
17	100	152	157	24
96	205	121	253	104

**3x3 Low-Pass Filter (Padding)**

57	91	96	95	60
86	125	127	126	88
60	93	107	103	70
76	110	141	122	88
46	77	110	90	60

You might recognize that a mean focal function has the property of smoothing out the variation in extreme values from cell-to-cell in the input raster. In fact, a mean focal function is special and also known as a **low-pass filter**, which has the effect of “blurring” a raster or image. On the other hand, we can apply a **high-pass filter** that performs edge enhancement. The kernel of a high-pass filter takes on a specific pattern of weights that usually involves a large weight to the focal (centre) cell of the kernel and negative weights to the neighbouring cells. Below is a common example of a high-pass kernel:

$$k = \begin{bmatrix} -0.7 & -1 & -0.7 \\ -1 & 6.8 & -1 \\ -0.7 & -1 & -0.7 \end{bmatrix}$$

The operation of a high-pass filter is to multiply the kernel weights above against the input raster cell values and then sum the result. The following is the result of applying the high-pass kernel weights above to our input raster with padding:

**Input Raster**

134	219	202	166	177
14	143	111	20	176
45	220	34	27	223
17	100	152	157	24
96	205	121	253	104

**3x3 High-Pass Filter (Padding)**

578	923	763	529	848
-534	118	-87	-789	642
-115	968	-573	-574	1165
-423	-121	128	274	-517
361	959	33	1215	320

## 9.9 Global

**Global functions** apply some operation to all cells in the raster. These are usually simply referred to as summary statistics of the raster since we usually want to know what the minimum, maximum, and average values are of all cells in a given raster. As a result, global functions do not return a raster as an output, but rather individual values, depending on the operation. Below are the summary statistics for our input raster:

Operation	Value
Minimum	14.00000
Maximum	253.00000
Mean	125.60000
Standard Deviation	75.53366

## 9.10 Zonal

Lastly, **zonal functions** perform an operation over some subset of cells defined by a zonal raster. Again, the operation can be any calculation of interest: mean, minimum, sum, etc. Zonal functions are useful for deriving information over different regions of a raster that share some thematic classification like land cover, ecosystem type, or jurisdiction. Suppose we have the following input raster and zonal raster:

**Input Raster**

134	219	202	166	177
14	143	111	20	176
45	220	34	27	223
17	100	152	157	24
96	205	121	253	104

**Zonal Raster**

1	1	1	1	1
1	1	1	2	2
2	2	2	2	2
2	2	3	3	3
3	3	3	3	3

Like a global function, zonal functions do not return an output raster, but rather individual values for each zone:

Zone	Minimum	Maximum	Mean	Standard.Deviation
1	14	219	145.75000	63.95032
2	17	223	95.77778	87.42394
3	24	253	139.00000	70.17122

## 9.11 Derivatives of Elevation Models

You can probably recognize by now that if we apply a raster function to a DEM, then there is a lot of derived information that we can exploit in a raster analysis of terrain. These are further classified as first order, second order, and compound terrain derivatives. In this next section we will explore several common derivatives of a DEM and, importantly, we will see how they appear differently in the flat terrain of Nelson River and the rugged terrain of Mount Assiniboine.

## 9.12 Slope

**Slope** is a first order terrain derivative using a focal raster function that represents the change in elevation over the distance of a cell edge. Slope can be expressed as a percentage or as angular degrees. As a percentage, a 0% slope would represent flat terrain and a 150% slope would represent steep terrain. It is worth pointing out that there is no upper bound for expressing slope as a percentage, but angular degrees are limited by an upper bound of 90°. It is mathematically impossible to achieve a slope of exactly 90° from a raster DEM because the cell size of a raster must always be great than 0 and a 90° angle would only be possible if two cells overlaid one another in the same raster, hence the impossibility. Slope is a derivative that is usually necessary in order to calculate other terrain derivatives.

Figure 9.3 shows the slope of Mount Assiniboine where red represents steeper slopes and green represents flatter slopes. For the Mount Assiniboine DEM, the slope values range from 2° at the flattest to 80° at the steepest. Can you see the peak yet?

Figure 9.4 shows the slope of Nelson River pouring into Hudson Bay. Again, red is steeper and green is flatter, but the slope values here only range from 0° to 9° and the vast majority of cells are less than 2°. Very flat terrain that can hardly be distinguished from the sea.

## 9.13 Aspect

**Aspect** is another first order terrain derivative that represents the azimuthal direction that a slope faces. For example, an azimuth of 0° is North-facing, 90° is East-facing, 180° is South-facing, and 270° is West-facing with all

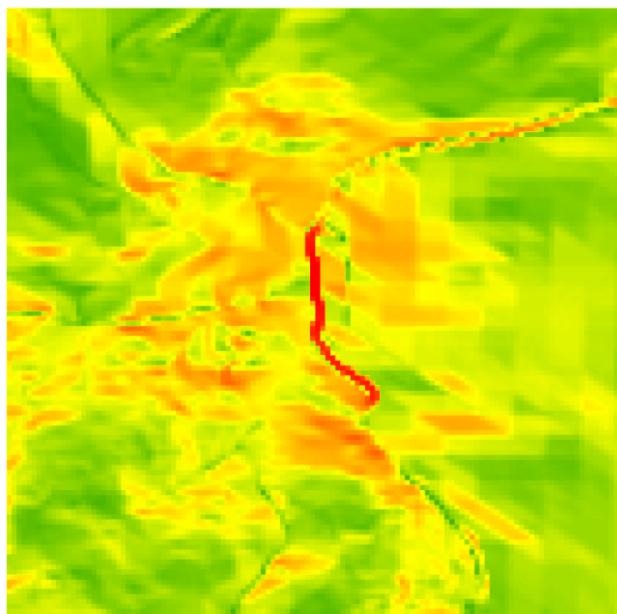


Figure 9.3: Slope of Mount Assiniboine at the border of Alberta and British Columbia, Canada. Data from Natural Resources Canada [2015]. Pickell, CC-BY-SA-4.0.



Figure 9.4: Slope of Nelson River pouring into Hudson Bay, Manitoba, Canada. Data from Earth Resources Observation and Science Center [2018]. Pickell, CC-BY-SA-4.0.

other azimuths in between. Typically, aspect is most apparent when azimuths are dominate throughout the DEM. For example, the slope aspects are clearly visible in Figure 9.5 showing Mount Assiniboine. By contrast, when slopes are very flat as is the case for the Nelson River, aspect can alternate frequently and produce a nearly random sequence except for the relatively steeper south bank of the river shown in Figure 9.6. It is possible to have an undefined aspect when the slope is  $0^\circ$ , which is the case for water seen coloured grey in Figure 9.6.

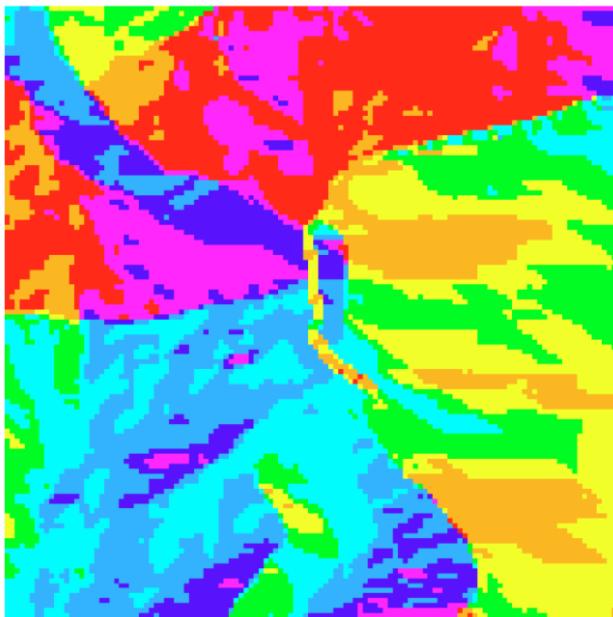


Figure 9.5: Aspect of Mount Assiniboine at the border of Alberta and British Columbia, Canada. Data from Natural Resources Canada [2015]. Pickell, CC-BY-SA-4.0.

## 9.14 Heat Load Index

One important derivative from aspect is a **heat load index (HLI)**, which quantifies the heat from incident solar radiation on a slope. There are several variations for computing the heat load on terrain, but a commonly used calculation is the HLI proposed by McCune and Keon [McCune and Keon, 2002]:

$$HLI = \frac{1 - \cos(aspect - 45)}{2}$$

If you imagine aspect represented by azimuths of a circle, then this calculation

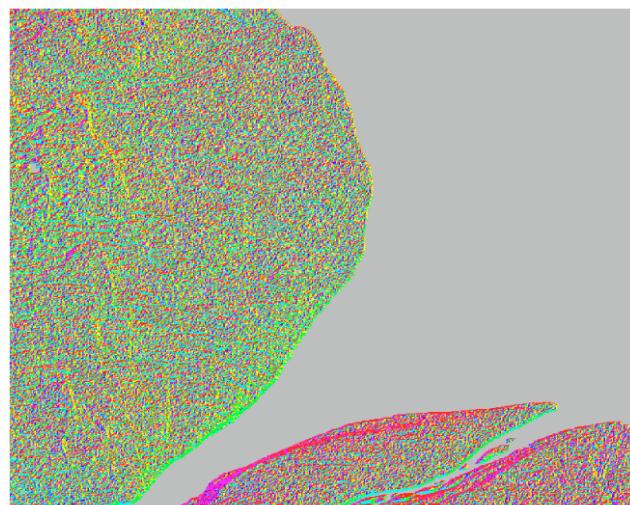


Figure 9.6: Aspect of Nelson River pouring into Hudson Bay, Manitoba, Canada. Data from Earth Resources Observation and Science Center [2018]. Pickell, CC-BY-SA-4.0.

has the effect of “folding” that circle in half along the  $45^\circ$  azimuth (northeast-southwest) so that southwest-facing slopes have higher values and northeast-facing slopes have lower values. Although both southeast- and southwest-facing slopes theoretically receive the same amount of incident solar radiation, this heat load index reflects the fact that southwest-facing slopes will be significantly hotter and drier, which can help inform vegetation potential and fuel moisture content.

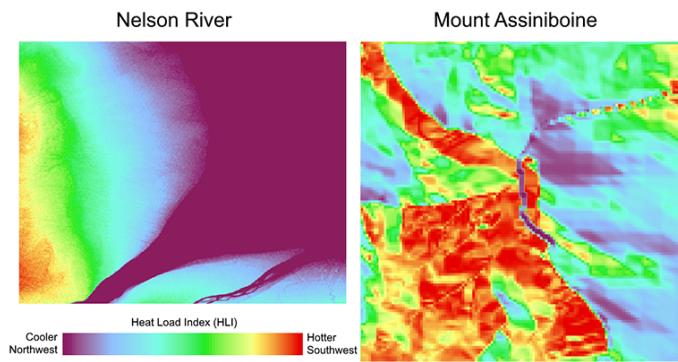


Figure 9.7: Comparing Heat Load Index (HLI) for Mount Assiniboine and the Nelson River. Pickell, CC-BY-SA-4.0.

## 9.15 Hillshade

A **hillshade** is a DEM with a simulated light source. The values in a hillshade represent the illumination of that cell given its slope, aspect, and position in the raster relative to a simulated light source. The light source has an elevation and azimuth that can be defined to reflect a particular time of day. Thus, hillshades can be useful to show a more 3-dimensional view of terrain from the bird’s-eye view of the DEM. Hillshades can also reveal fine differences in terrain that are not apparent from a simple DEM. Figure 9.8 animates (online) a series of hillshades created with azimuths at  $10^\circ$  intervals from  $0^\circ$  to  $350^\circ$  with a light source at an elevation of  $45^\circ$  for Mount Assiniboine. This has the effect of rotating an approximately 9 AM high Sun on the solstice around Mount Assiniboine.

A hillshade is still a 2D raster with 3D features simulated through an illumination process. A hillshade does not reveal much for flatter terrain since there is not much variation in slopes across the entire raster, as can be seen in Figure 9.9 for Nelson River, which is also animated in the same way as Figure 9.8. As you can see, there is no change over the water of the river and Hudson Bay where slope is  $0^\circ$  and aspect is undefined and only some minor terrain features are apparent over the land. Still, if you watch closely, some finer scale features

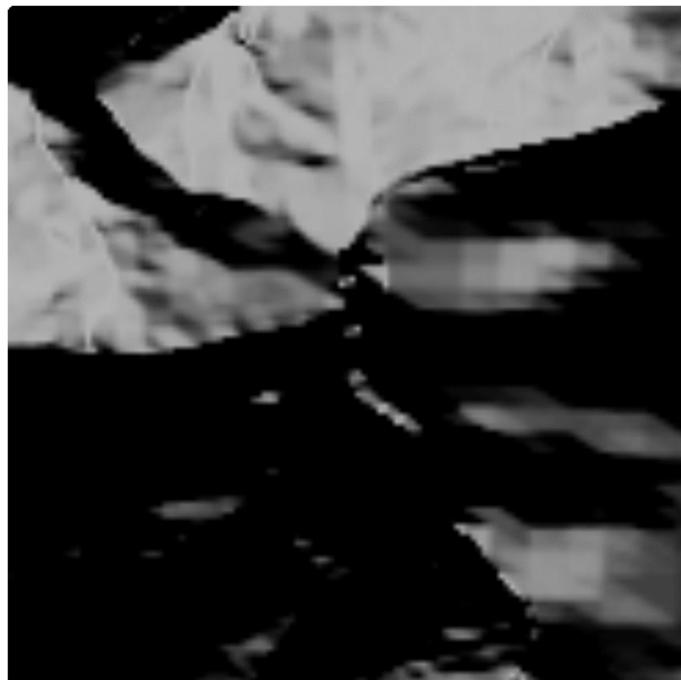


Figure 9.8: Hillshade of Mount Assiniboine at the border of Alberta and British Columbia, Canada. Data from Natural Resources Canada [2015]. Online version is animated across all azimuths from  $0^\circ$  to  $350^\circ$  by  $10^\circ$  intervals. Pickell, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengis.ca/raster-analysis-and-terrain-modelling.html#fig:9-mount-assiniboine-hillshade-animation>.

can be made apparent that are not otherwise visible from the original DEM and other terrain derivatives.

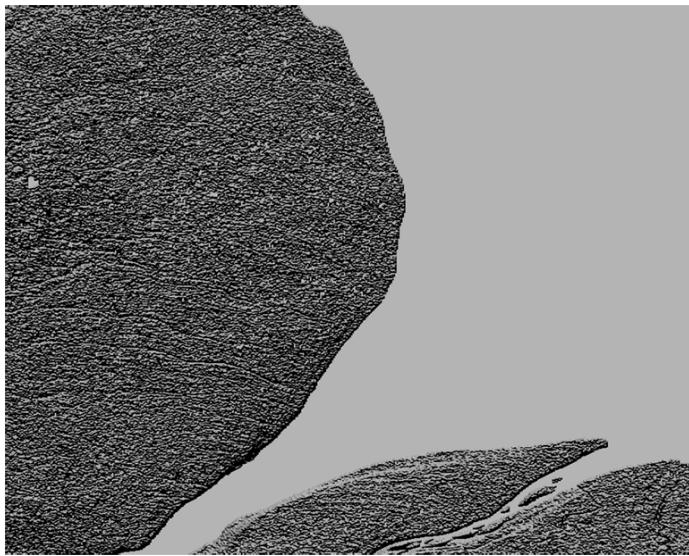


Figure 9.9: Hillshade of Nelson River pouring into Hudson Bay, Manitoba, Canada. Data from Earth Resources Observation and Science Center [2018]. Online version is animated across azimuths from  $0^\circ$  to  $350^\circ$  by  $10^\circ$  intervals. Pickell, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/raster-analysis-and-terrain-modelling.html#fig:9-nelson-river-hillshade-animation>.

## 9.16 Sinks, Peaks, and Saddles Oh My!

**Sinks** are depressions in a DEM where the focal cell is surrounded by eight neighbouring cells with higher elevations (Figure 9.10). Sinks are often naturally occurring (e.g., lakes, ponds, and wetlands), but can also be due to random error, elevation precision, DEM cell size, or other pre-processing that may have been applied to the DEM such as mosaicking. The problem with sinks when modeling runoff is that water will enter the cell, but will not be able to exit in any direction, so they must be filled prior to using a DEM in any hydrology workflow.

**Peaks** are the opposite of sinks, where a focal cell is surrounded by eight neighbouring cells that share the same lower elevation (Figure 9.11). Like sinks, peaks are also naturally occurring (e.g., mountain peaks and ridge lines) and may also be artifacts of data resolution and processing.

**Saddles** occur when a lowland is bounded by two or more peaks (Figure 9.12). Usually, saddles identify a divide between drainage basins because they often

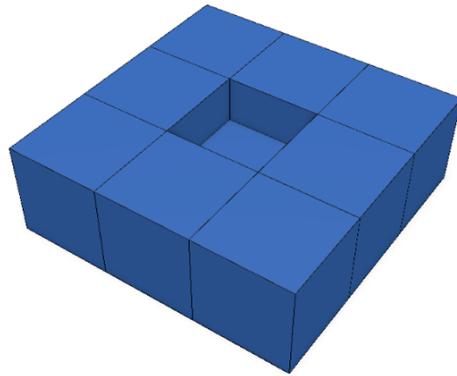


Figure 9.10: A sink is shown in 3D for an array of 3x3 pixels. Pickell, CC-BY-SA-4.0.

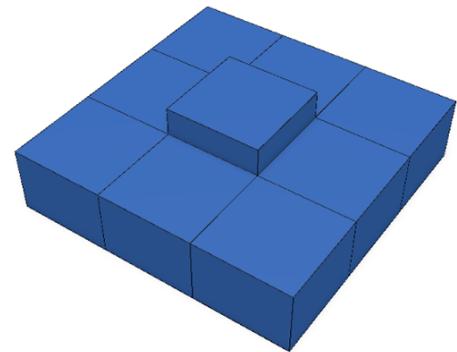


Figure 9.11: A peak is shown in 3D for an array of 3x3 pixels. Pickell, CC-BY-SA-4.0.

occur along ridgelines and precipitation will runoff into one or the other drainage basin, but not both.

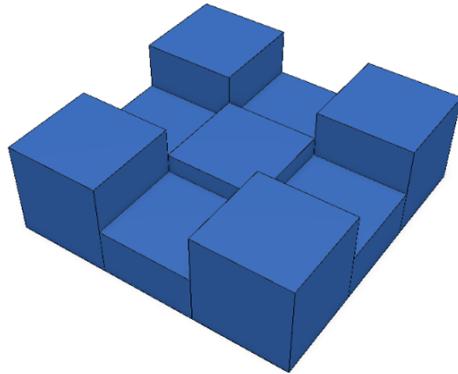


Figure 9.12: A saddle is shown in 3D for an array of 3x3 pixels. Pickell, CC-BY-SA-4.0.

## 9.17 Landform Classification

One of the motivations for deriving information from a DEM is to classify terrestrial landforms. Some of the derived information we have seen so far, like slope, can be used to classify landforms, but slope does not inform on geographic position. For example, a low slope value could be a plateau on top of a mountain or a river in a valley bottom. Therefore, we need more contextual information beyond these first or second terrain derivatives that we have looked at so far. There are two popular ways to derive these complex terrain derivatives by calculating terrain curvature and topographic position indices, which is what we will explore next.

## 9.18 Profile and Planform Curvature

Curvature of terrain is described as convex, flat, or concave, which impacts the flow acceleration or deceleration of runoff over the terrain. There are two components of terrain curvature representing the x- and y-axes of geographic space. **Profile curvature** describes the downslope curvature while **plan or planform**

**curvature** describes the curvature that is perpendicular to the downslope curvature. In this way, the vertical expression of terrain can be described by two, perpendicular curves. Figure 9.13 illustrates the nine different combinations of profile and planform curvature.

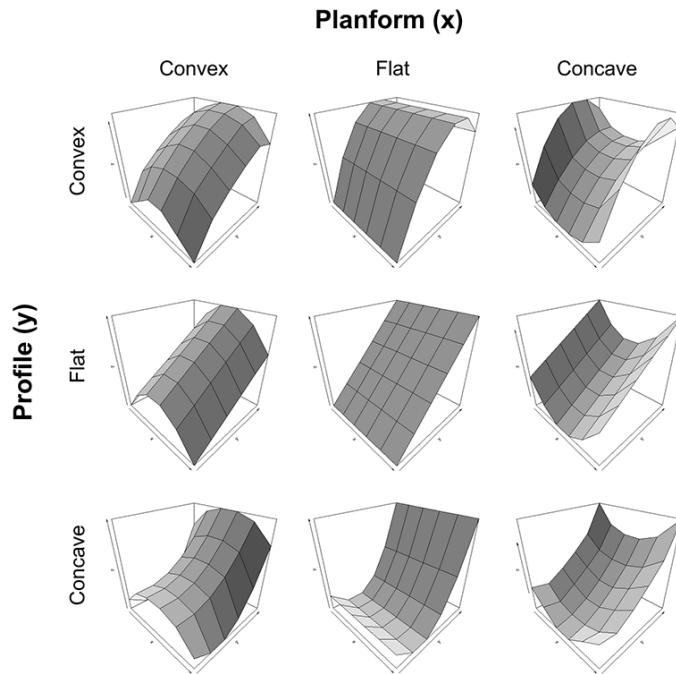


Figure 9.13: Conceptual diagram showing surfaces of all combinations of profile and planform curvature. Pickell, CC-BY-SA-4.0.

There are many ways to calculate curvature, which usually take the form of a polynomial that is fit by a focal function over some neighbourhood of pixels. A 3x3 kernel is typically used to calculate profile and planform curvature, however, it is also common to increase the kernel size to counteract the effect of noise in high resolution DEMs. Figure 9.14 illustrates examples of 3x3 kernels for each of the profile and planform curvature combinations. You might recognize variations of sinks, peaks, and saddles amongst the examples, which are of course not exhaustive of all possible elevation value combinations, but they are otherwise representative of the primary patterns that are indicated by profile and planform curvature.

Both profile and planform curvature are calculated as continuous floating point integers where: positive values represent convex curvature in the profile and concave curvature in the planform; zero represents no curvature; and negative values represent concave curvature in the profile and convex curvature in the planform. Figure 9.15 shows the profile and planform curvature for Mount

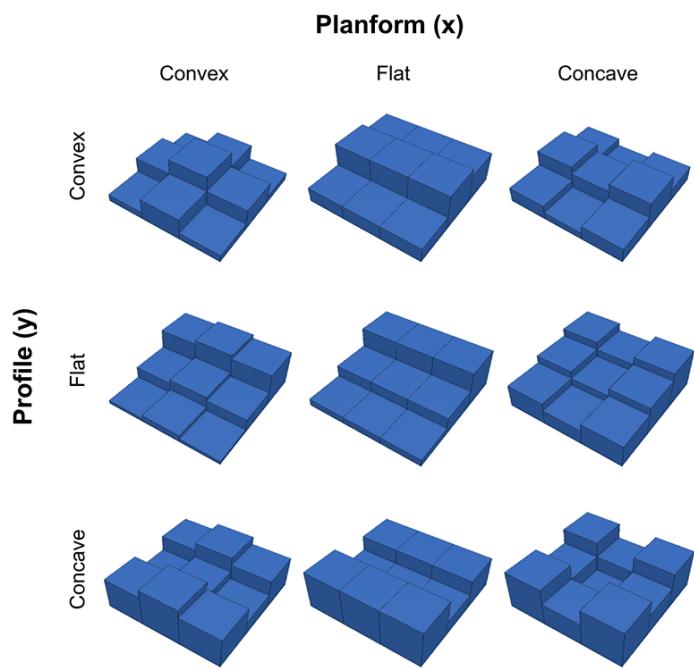


Figure 9.14: Conceptual diagram showing 3x3 kernel examples of all combinations of profile and planform curvature. Pickell, CC-BY-SA-4.0.

Assiniboine. What landforms can you start to identify from terrain curvature? Can you identify any sinks, peaks, or saddles in the DEM?

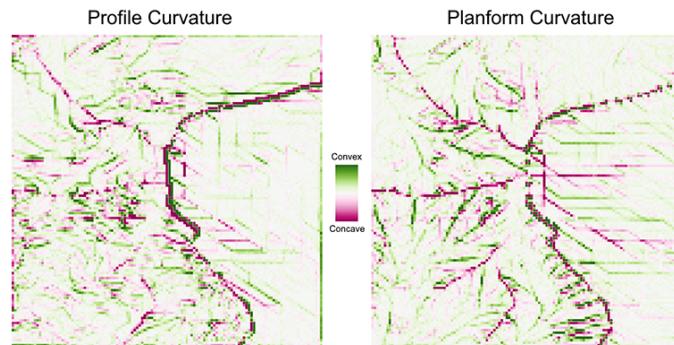


Figure 9.15: Profile and planform curvature for Mount Assiniboine at the border of Alberta and British Columbia, Canada. Data from Natural Resources Canada [2015]. Pickell, CC-BY-SA-4.0.

## 9.19 Topographic Position Index

Another way to derive and classify landforms is through the use of a **topographic position index (TPI)**, which is a focal function that accounts for the difference between the focal cell elevation and the average elevation of the eight neighbouring cells (in the case of a 3x3 kernel). Therefore, positive TPI values indicate ridges or peaks, zero indicates either a constant slope or a flat area or a saddle, and negative TPI values indicate a valley or sink. Figure 9.16 shows the TPI for Mount Assiniboine compared with the Nelson River. The TPI is sometimes also referred to as a terrain ruggedness index and there are other variations for calculating it as well (e.g., computing the standard score of the focal cell instead of the mean difference).

## 9.20 Hydrology Work”flows”

By now, you should recognize that water can be modelled over a DEM by making the simple assumption that water will flow from higher elevations to lower elevations. We can undertake this modelling using a DEM because the continuous surface of elevation values can represent the runoff process. In this next section, we will look at how we can extract everything from runoff accumulation to stream networks and watersheds from a DEM.

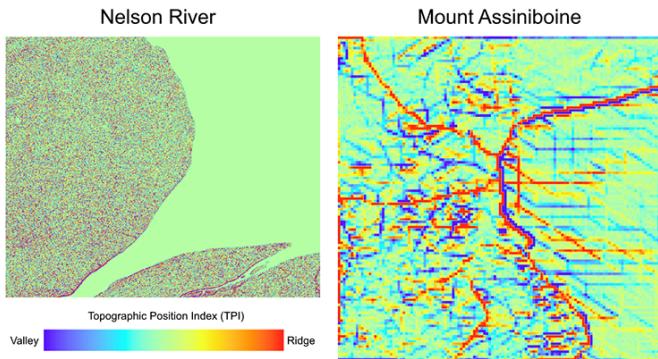


Figure 9.16: Comparing Topographic Position Index (TPI) for Mount Assiniboine and the Nelson River. Data from Natural Resources Canada [2015] and MAXAR, respectively. Pickell, CC-BY-SA-4.0.

## 9.21 Flow Direction and Flow Accumulation

Runoff is the most fundamental process that can be modelled from a DEM. In fact, modelling runoff supports nearly every subsequent derived hydrological variable of interest (e.g., flow accumulation, stream order, and watersheds). The principle assumption is that water flows from higher elevations to lower elevations. So the first step in any hydrological workflow is to calculate **flow direction** or the path that water will flow given precipitation over a focal cell in the DEM. This is made possible because we also know the eight neighbouring cell elevations. There are several algorithms for calculating and assigning flow direction, but the simplest is a focal function known as *D8*, first proposed by Greenlee [1987], that answers the question, *which neighbouring cell has the lowest elevation?* With only eight neighbours, runoff can only travel in eight unique azimuths. Flow direction can be encoded as an azimuth with valid values of  $FlowDir \in \{0, 45, 90, 135, 180, 225, 270, 315, 360\}$  (16-bit integer) where 0 is reserved for undefined flow direction, but more D8 flow direction is commonly encoded as an 8-bit integer with valid values of  $FlowDir \in \{1, 2, 4, 8, 16, 32, 64, 128, 255\}$  where 255 is reserved for undefined flow direction. Once we know the flow direction we can calculate the **flow accumulation**, which is a focal function tally of the number of upslope DEM cells that flow into the focal cell. Figure 9.17 illustrates how flow direction and flow accumulation work together to model runoff over a DEM surface.

Trouble arises when there are two or more neighboring cells with the same minima (e.g., a peak or saddle). This situation can arise for very flat terrain where the elevation might be the same in nearly every direction. As we have already seen, sinks can cause water to flow into but not out of the focal cell, and therefore result in an undefined flow direction. Sinks must therefore be identified and filled in a DEM before calculating flow direction. There are several algorithms

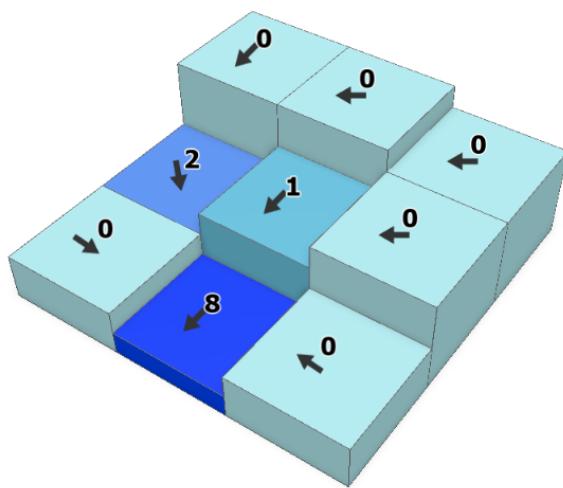


Figure 9.17: Flow direction shown in arrows and flow accumulation of upslope cells shown in numbers for the D8 flow direction algorithm. Pickell, CC-BY-SA 4.0.

for filling sinks, but generally the process involves calculating flow direction and then moving downslope from the highest elevations of the DEM to the valley bottoms and identifying depressions, which are then filled. This process continues to iterate until no more depressions are identified [Marks et al., 1984]. Figure 9.18 animates this iterative process of filling sinks in a cross-sectional profile of elevation, but the process is actually applied simultaneously over both geographic dimensions. Each iteration can introduce new sinks since the filling process directly modifies the DEM, so flow direction must be recalculated after each iteration in order to solve the next set of sinks. As a result, this filling process can be quite time consuming and usually stopping criteria are imposed such as a minimum depth threshold or a minimum slope for all cells in the DEM (i.e., no cells with flat slopes) to ensure that runoff is continuously downslope.

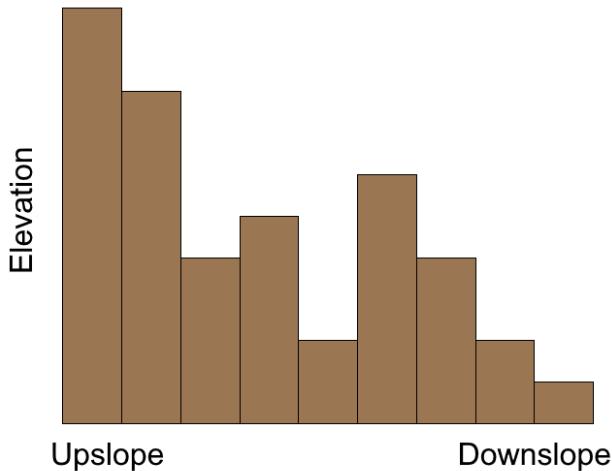


Figure 9.18: Animated process of filling sinks in a DEM. Pickell, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/raster-analysis-and-terrain-modelling.html#fig:9-sink-fill>.

The D8 algorithm is intuitive, fast, simple to implement and you will probably see it as the default option across many different GIS software. However, it is very sensitive to sinks, flat areas, and otherwise erroneous pixels because the D8 algorithm forces flow to a single raster pixel. If the raster pixel receiving the upslope flow is erroneous, then all downslope flow direction and, by extension, flow accumulation may also be erroneous. The D8 algorithm was developed in the mid-1980s during a time when computers such as the AT&T 3B2/300 “mini computer” shipped with a 30 Mb hard disk drive, 1 Mb of random ac-

cess memory, and a 10 Mhz central processing unit, all of which would have cost about US\$10,000 in 1984. To put that in perspective, the 16-bit DEM of the Nelson River in Figure 9.1 is 1.3 Mb and the combination of all the terrain derivatives that we have covered so far in this chapter (slope, aspect, HLI, hillshade, TPI, profile curvature, planform curvature, flow direction and flow accumulation) amounts to 19.5 Mb of disk space for a single, modest-sized study area (938 x 751 pixels).

More sophisticated flow direction algorithms have developed since then that are more robust against erroneous pixels in the input DEM, including the D-infinity flow (DINF) algorithm proposed by Tarboton [Tarboton, 1997] and the multiple flow direction (MFD) algorithm developed by Qin et al. [Qin et al., 2007]. The DIF algorithm works by creating triangular facets within the 3x3 kernel, then flow direction is assigned as the azimuth of facet with the steepest slope. Calculated in this way, flow direction is represented as a continuous value between 1-360 instead of one of eight possible values with the D8 method. The MFD algorithm works by partitioning downslope flow to all eight neighbours. As a result, each neighbour receives a fraction of flow and the flow accumulation function then accumulates these fractions downslope. Figure 9.19 shows the D8 flow direction and associated flow accumulation for Mount Assiniboine. Once flow direction and accumulation rasters have been generated, the floodgates open for a wide range of other derivative calculations, which are discussed in the next sections.

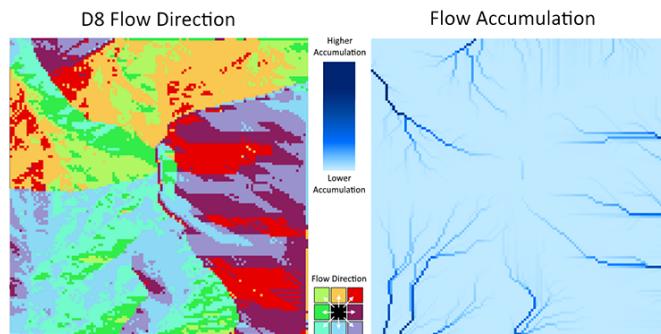


Figure 9.19: D8 Flow direction and flow accumulation for Mount Assiniboine at the border of Alberta and British Columbia, Canada. Data from Natural Resources Canada [2015]. Pickell, CC-BY-SA-4.0.

## 9.22 Stream Delineation

Streams and rivers represent paths of accumulated flow and can be modeled from a flow accumulation raster by applying a simple threshold: values above

this threshold are considered rivers or streams. The threshold that is applied to a flow accumulation raster should be selected based on expert knowledge and visual interpretation of the landscape. The actual value of the flow accumulation threshold is directly related to the extent of the flow accumulation raster. That is, larger flow accumulation rasters have more cells and therefore the accumulative values will be larger compared with a smaller flow accumulation raster. This means that a flow accumulation threshold of 100 can represent the same river that might be represented by a flow accumulation value of 10,000 in a larger raster. Thus, some initial visualization of the flow accumulation raster like in Figure 9.19 is needed to determine what threshold should be applied. Once a threshold is determined, the flow accumulation raster can be reclassified to a binary (1 or 0) raster where 1 represents the presence of a stream or river channel. Usually, streams and rivers are represented using a vector data model instead of a raster, so the final step is to convert the binary raster to a set of line features that can be used in other workflows like network analysis (see Chapter 8).

## 9.23 Stream Order

Once a stream raster has been delineated from a flow accumulation raster, then it is possible to derive stream order. **Stream order** is an ordinal value that describes the level or hierarchy of branching in a stream network. Higher stream order values generally represent larger streams that are closer to an output such as the ocean. Thus, smaller stream orders will generally be found “upstream” at higher elevations. There are two primary methods for assigning a stream order value. Both methods assign the value of 1 to segments of stream networks that are the “outermost” branches or the generally the highest in elevation. Thus, a first order stream would have relatively smaller flow accumulation values than a second order stream because it drains a smaller area. The way the Strahler [Strahler, 1957] and Shreve [Shreve, 1966] methods differ is how higher order stream values are assigned when two stream segments merge. For both methods, if two first (1-)order streams merge, then the next downstream segment is assigned a value of 2. However, if a first (1-)order and second (2-)order stream merge, then the Strahler method assigns the next downstream segment the higher of the two orders, which is 2 in this case, while the Shreve method adds together the orders of the upstream segments to assign the new value, which is 3 in this case. This pattern is repeated until all downstream segments have been labelled.

In the GIS, stream order can be automatically calculated by using the binary stream delineation raster (1 = stream) and the flow direction raster. In other words, the GIS tool needs to know what the stream skeleton is and in which direction it is flowing. The output is usually a raster that contains the stream order values for the streams delineated in the binary stream raster. Again, the stream orders can be easily converted to line features for further vector-based

analysis.

## 9.24 Flow Length

**Flow length** describes the distance of a stream path along the flow direction raster. This is calculated for all cells in the flow direction raster, not just the stream segments. In other words, each cell in the output raster takes on the value of how far water would flow downstream. Since the flow length of all cells in the output raster are known, it is possible to derive the longest flow path within a given basin. The longest flow path can be used to describe the time of concentration within a basin, which is a measure of the amount of time (i.e., a function of distance) that a precipitation event would take to exit or “flush” out of a given basin at an outlet. With some GIS software, flow length may be one of many optional outputs when calculating flow direction and flow accumulation from a DEM. You may also have the option of applying a raster of flow barriers, which modify the flow path and therefore the flow length.

## 9.25 Watershed Delineation

**Watersheds** represent a contributing area to a particular pour point and therefore the total upslope area that flows or drains through a particular cell of the DEM. **Pour points** represent the constrained location that upslope flow must pass through and is identified as a single cell in a raster. Thus, if a set of pour points are defined, then it is possible to simply look at the neighbours of any pour point and identify which of the neighbouring cells flow into the pour point cell. If a neighbouring cell flows into the pour point, then that neighbouring cell is labelled within the same watershed as the pour point. This iterative process continues until the edge of the watershed search boundary finds a ridge, in which case the flow direction of neighbouring cells would flow opposite to the watershed and the search would end in the iteration that adds no more cells to the watershed. This process is repeated for all pour points until all cells in the DEM are classified into their respective watersheds.

You might be wondering how pour points are initialized for this process. A common way to initialize pour points is to manually identify them within the context of your research objectives. Practically, pour points should be located on cells with high flow accumulation (i.e., not a random slope), so that you can be sure your watersheds will reach from ridge-to-ridge. Pour points can also be programmatically identified in several ways. For one, you could select all nodes of your stream network where the stream order transitions from a specific lower order to the next highest order number. For example, setting pour points to the transitions from third order to fourth order streams would yield a set of watersheds that contribute to fourth order streams with the caveat that you do not identify any watersheds downstream of fourth order streams. You can adjust this approach as needed if you are looking at larger drainage areas by

increasing the stream order that you are considering. Typically, the transition between your highest order stream and the next lowest order represents all of the drainage areas for your DEM. If you need regularly-sized watersheds, then the flow accumulation raster conveniently indicates how many upslope cells flow into each downslope cell, so you can identify pour points along your stream network that contribute exactly 10,000 upslope cells. If your DEM cell size is 10 m, then this would be equivalent to mapping 1 km<sup>2</sup> watersheds. If your goal is regularly-distributed watersheds along a particular reach of your stream network, then you could export the target stream segment and create points that are spaced at regular intervals along that stream segment. More on network analysis of hydrological networks in Chapter 8.

## 9.26 Topographic Wetness Index

**Topographic Wetness Index (TWI)** is a simple calculation that can be used to identify locations where draining water is likely to collect or pool. Sometimes also referred to as the Compound Topographic Index (CTI), the calculation is as follows:

$$TWI = \ln \frac{a}{\tan(b)}$$

where  $a$  is the local upslope area for a given cell and  $b$  is the local slope (radians) for a given cell. This calculation is essentially a ratio of flow entering a cell and the discharge of that flow out of the cell, represented by the tangent of the slope. Figure 9.20 illustrates an overview of the components of the TWI calculation in a conceptual geographic space.

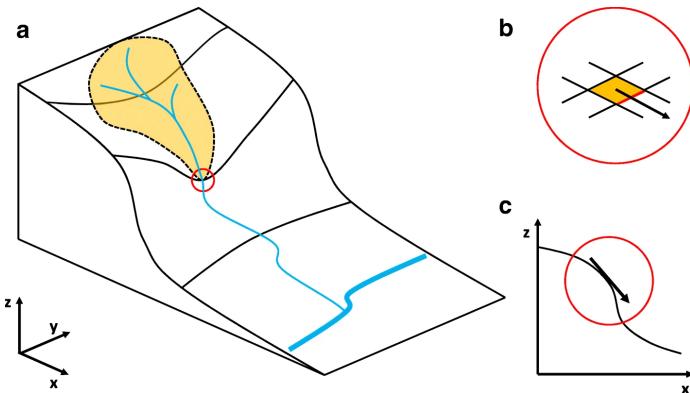


Figure 9.20: Conceptual figure showing how Topographic Wetness Index (TWI) is a function of (a) upslope area, (b) area and direction of local flow, and (c) the tangent of the local slope angle. Reproduced from Mattivi et al. [2019], CC-BY-4.0.

The result is that cells with high flow accumulation and gentle slopes will have high TWI values (water pools), while cells with low accumulation and steep slopes will have low TWI values (water flows). The actual values of this ratio are unitless, but they are relative. If you are using the flow direction raster directly in your calculation, it is important to multiply  $a$  by the area represented by each cell. For example, a 10 m resolution DEM would use the following calculation:

$$TWI = \ln \frac{a \times 100}{\tan(b)}$$

## 9.27 Case Study: Topographic Indices for Wetland Mapping

*Case Study Author: Ramon Melser (CC BY 4.0. unless otherwise indicated), University of British Columbia, Master of Geomatics for Environmental Management graduate, 2021*

Natural processes on a landscape can be interpreted by analyzing its topography. DEMs can be used to create a wide range of topographic indices, which may be used to inform on landscape geomorphology, hydrology and biological processes in a study area [Mattivi et al., 2019]. To demonstrate the applications of topographic indices, and attempt was made to predict the distribution of wetlands across a study area in the Fort St. John Timber Supply Area, in North-eastern British Columbia, Canada (Figure 9.21).

The use of topographic indices was employed alongside spectral indices to perform land cover classification. For the successful prediction of wetland landscape classes, a particularly promising topographic index is Topographic Wetness Index (TWI). In brief, high TWI values indicate areas with high water accumulation and low slope, whilst low TWI values indicate areas that are well drained and on steep slopes. There are several GIS platforms that provide users with TWI tools, yet not all are the same. A particular tool used in this case study was the SAGA Wetness Index tool: relative to ‘regular’ TWI calculations, System for Automated Geoscientific Analyses (SAGA) Wetness modifies the calculation of the contributing catchment area to a cells ‘wetness’, and consequently produces a more realistic prediction of soil moisture potential [Mattivi et al., 2019]. The equations for SAGA Wetness ( $WI_S$ ) (Equation 1), and for its unique computation of catchment area ( $SCA_M$ ) (Equation 2) are given below [Böhner and Selige, 2006].

$$WI_S = \ln\left(\frac{SCA_M}{\tan\beta}\right) \quad (\text{Equation 1})$$

$$SCA_M = SCA_{max} \left(\frac{1}{15}\right)^{\beta \exp(15^\beta)} \quad (\text{Equation 2})$$

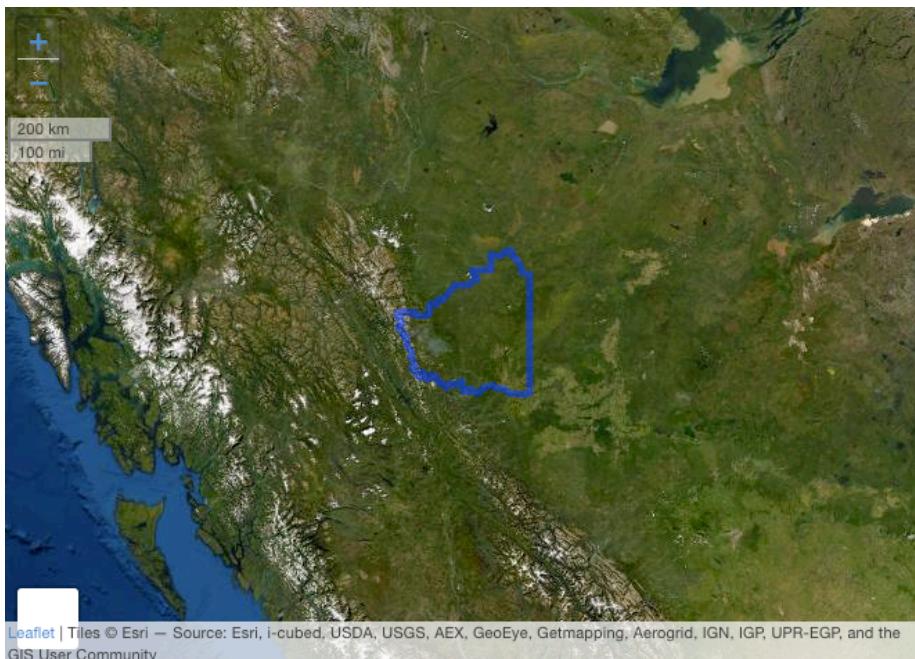


Figure 9.21: The Fort St. John Timber Supply Area, British Columbia, Canada. Data from Government of British Columbia and licensed under the Open Government Licence - British Columbia. Mesler, CC-BY-SA-4.0. Interactive figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/raster-analysis-and-terrain-modelling.html#fig:9-fort-st-john-study-area>

Both are functions of slope angle  $\beta$ , and  $SCA_{max}$  is defined as drainage area per unit contour width ( $m^2*m^{-1}$ ). SAGA Wetness across the study site is shown in Figure 9.22.

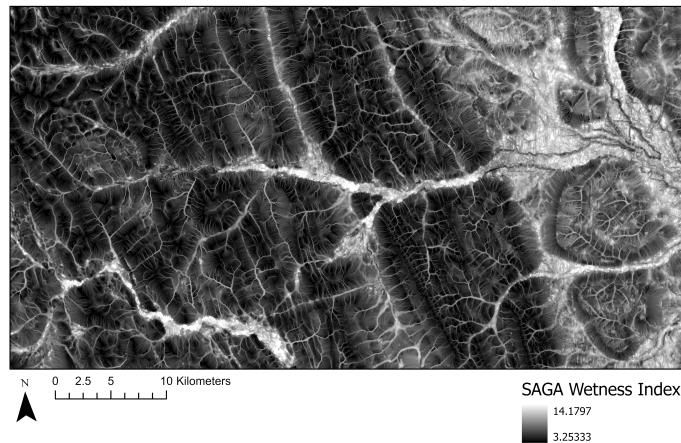


Figure 9.22: System for Automated Geoscientific Analyses (SAGA) Wetness Index across the Chowade Watershed. Melser, CC-BY-SA-4.0.

In addition to SAGA Wetness, another useful index that was included in the Case Study was the Multi-Resolution Valley Bottom Flatness (MRVBF) index. Introduced by [Gallant and Dowling, 2003], this index identifies hill slopes and valley bottoms, and characterizes hydrological and geomorphic dynamics. This index serves to identify areas that are likely to be occupied by water, and has been used widely in remote sensing studies [Huang et al., 2017]. MRVBF is derived from flatness and lowness characteristics and is computed using slope and ranked elevation, see Figure 9.27 [Gallant and Dowling, 2003]. Note that this is a simplification of the computational process, and Gallant and Dowling may be consulted for a more in-depth breakdown of the 18 equations involved in computing MRVBF. Figure 9.23 shows MRVBF across the study site.

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is
```

Multi-Resolution Valley Bottom Flatness (MRVBF) Computational Components described by Gallant and Dowling [2003]. Melser, CC BY 4.0.

Figures 9.24 and 9.25 below demonstrate the relationship between Elevation and MRVBF through a comparison of cross sections from both layers. As is shown by the cross-sections of MRVBF and the DEM, high MRVBF values are directly correlated with low slope and elevation areas in the DEM.

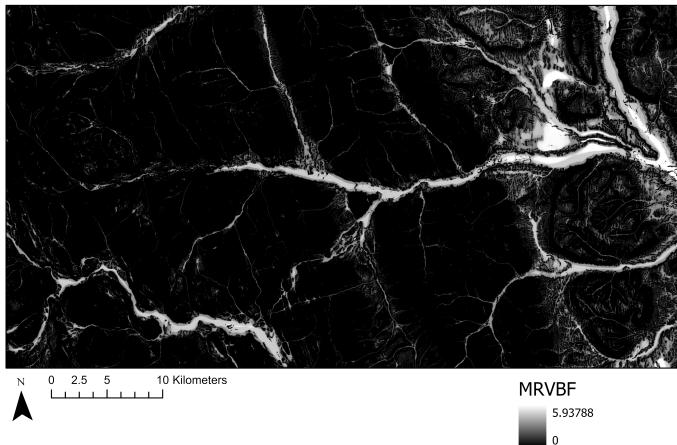


Figure 9.23: Multi-resolution Valley Bottom Flatness (MRVBF) across the Chowade Watershed. Melser, CC-BY-SA-4.0.

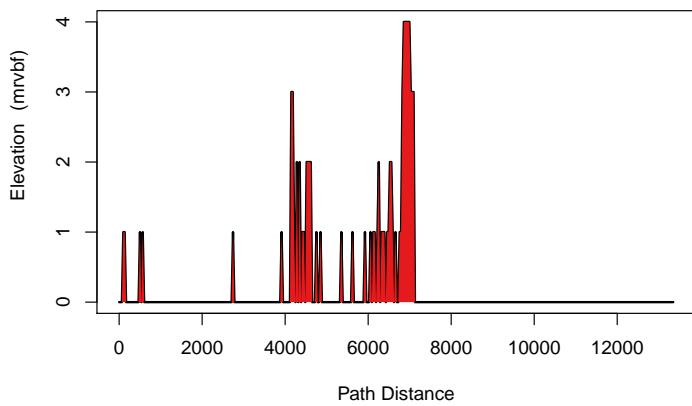


Figure 9.24: Cross Section where elevation is expressed through Multi-Resolution Valley Bottom Flatness (MRVBF). High values represent valley bottoms, and low values represent steep slopes. Melser, CC-BY-SA-4.0.

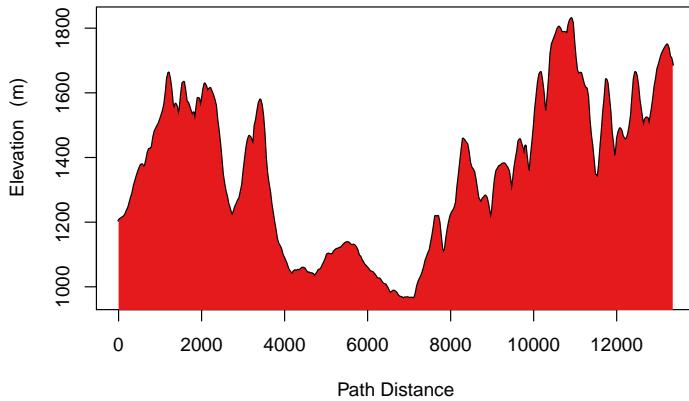


Figure 9.25: Elevation Cross Section. Melser, CC-BY-SA-4.0.

## 9.28 DEM Derivatives and Classification

To exemplify the potential of topographic indices, SAGA Wetness, MRVBF, Slope and Aspect are combined with spectral information from satellite imagery in a classification algorithm. Using these layers, predictive mapping of three simple land cover classes - Alpine, Forested, and Wetland - was performed in the Chowade Watershed Management basin [City of Vancouver, 2009]. The interactive map below in Figure ?? allows for navigation of the classified map, as well as the input layers. An exploration of the input layers and prediction raster make it apparent how certain values in the input layers informed the prediction of wetlands. In other words, low slope values, high SAGA Wetness values and valley bottoms in MRVBF. The overall accuracy from this particular classification is 84.4%, showing the value of layer derivatives in classification exercises.

## 9.29 3D Geovisualization

**Geovisualization** is simply the process of visualizing geographic information at any stage of the GIS workflow. 3D geovisualization is distinguished by the motivation to create immersive scenes that utilize an elevation, height, or z-profile. Like their 2D map counterparts, 3D geovisualizations should be thought of as spatial models. 3D geovisualizations can represent real phenomena such as landforms and buildings, but they may also be non-real representations that map some phenomenon in the traditional x and y dimensions and mapping an attribute to the z dimension that is not necessarily height or elevation. The results can be dramatic forms to convey geospatial information, sometimes requiring specialized equipment to view them such as virtual reality headsets, polarized or shutter glasses, red/cyan glasses, and 3D stereo monitors. We will focus on a few digital options in this section.

## 9.30 Anaglyphs

**Anaglyphs** are composed of two images that are simultaneously displayed in two colours, typically red and cyan. The result when viewed through red/cyan filtered lenses is a stereographic view of the frame. Anaglyphs are among the simplest 3D geovisualizations that are frequently used to represent terrain and landforms as the effect is essentially adding depth to an otherwise 2D frame. The depth of terrain is best visualized obliquely with an anaglyph as in Figure 9.26 below showing Mount Odin, the highest peak on Baffin Island in Nunavut, Canada.

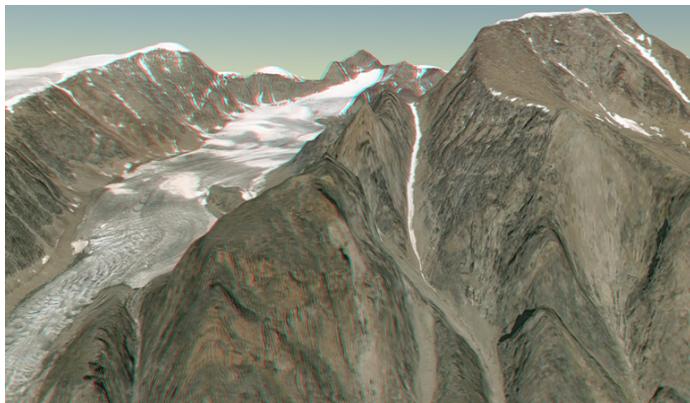


Figure 9.26: Anaglyph of Mount Odin on Baffin Island, Nunavut, Canada. Imagery by MAXAR. Pickell, CC-BY-SA-4.0.

## 9.31 Viewsheds

A **viewshed** is an analysis that simulates a physical line-of-sight using a DEM to identify pixels that can be seen from an observer point or set of points. The result of a viewshed is usually a raster with the same dimensions as an input DEM that has pixels encoded with values of 1 indicating the pixel can be seen from the observer point. In other words, the elevations for each pixel in the DEM is used to calculate whether a line-of-sight would pass through it given some other observer location at another elevation. Viewsheds can be useful for incorporating landscape aesthetics into GIS workflows where an area may be visually sensitive. For example, forest harvests can be designed and planned to mitigate the visual impact at nearby towns or highways. As well, observing pull-outs along scenic highways can be identified using viewsheds.

## 9.32 Extrusion

**Extrusion** is the process of taking 2-dimensional features such as points, lines, or polygons and assigning an elevation to each feature so that it can be visualized in 3-dimensions. An extrusion is often performed using the values of an attribute in a feature class. In this way, extrusion can be used to create 3D buildings from 2D polygons of building footprints that contain an attribute of height (Figure 9.27).

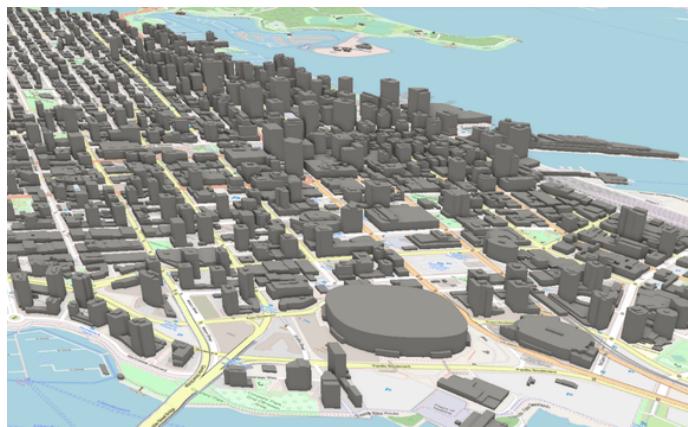


Figure 9.27: Animation showing building polygons in Vancouver, British Columbia extruded by a height attribute. Building data from City of Vancouver [2009], licensed under Open Government License - Vancouver. Base map © OpenStreetMap contributors, licensed under Open Data Commons Open Database License. Pickell, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/raster-analysis-and-terrain-modelling.html#fig:9-vancouver-building-height-extruded>

Extrusion is also frequently used to visualize non-height attributes. For example, population density can be extruded as a “height” value to give a unique perspective of the relative change in this attribute across space that may otherwise be difficult to appreciate from a 2D map with colour alone (Figure 9.28).

## 9.33 Exaggeration

**Exaggeration** is the process of multiplying real height or elevation values by a constant factor in order to create a larger range of values so that vertical features and patterns are made more apparent. Figure 9.29 illustrates an example of exaggeration of a DEM for Mount Logan, the highest peak in Canada (5,959 m), located in the Yukon Territory.

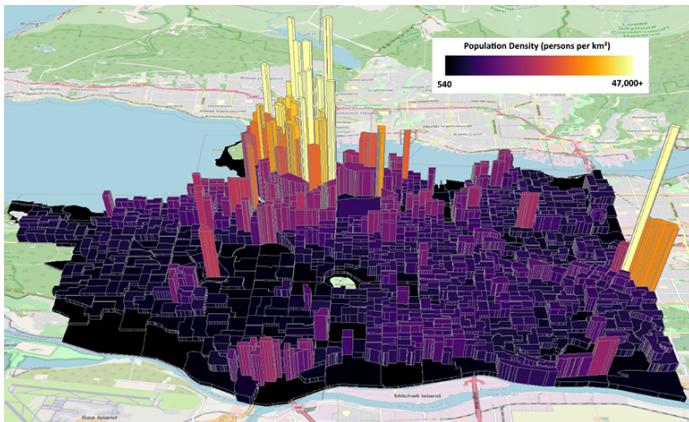


Figure 9.28: Census dissemination areas extruded by population density (2016) for Vancouver, British Columbia. Census data from Government of Canada [2017], licensed under Statistics Canada Open License. Reproduced and distributed on an ‘as is’ basis with the permission of Statistics Canada. Base map © OpenStreetMap contributors, licensed under Open Data Commons Open Database License. Pickell, CC-BY-SA-4.0.

## 9.34 Summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut in dolor nibh. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent et augue scelerisque, consectetur lorem eu, auctor lacus. Fusce metus leo, aliquet at velit eu, aliquam vehicula lacus. Donec libero mauris, pharetra sed tristique eu, gravida ac ex. Phasellus quis lectus lacus. Vivamus gravida eu nibh ac malesuada. Integer in libero pellentesque, tincidunt urna sed, feugiat risus. Sed at viverra magna. Sed sed neque sed purus malesuada auctor quis massa.

## Reflection Questions

1. Explain ipsum lorem.
2. Define ipsum lorem.
3. What is the role of ipsum lorem?
4. How does ipsum lorem work?

## Practice Questions

2. Given ipsum, solve for lorem.
3. Draw ipsum lorem.

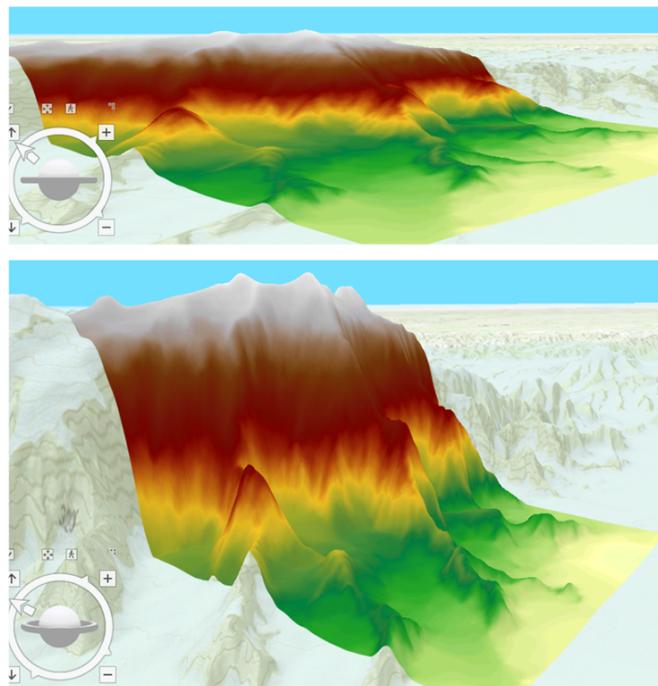


Figure 9.29: Elevation exaggeration of Mount Logan elevation, Yukon Terrotiry, Canada. Digital Elevation Model by Natural Resources Canada [2015]. Pickell, CC-BY-SA-4.0.

## **Recommended Readings**



## Chapter 10

# Spatial Estimation

Written by Sushil Nepal and Paul Pickell

**Geostatistics** uses the metrics based on statistical tools that are used to characterize the distribution of an event across the geographical region of interest [Getis et al., 2004]. In context of spatial data, it is important to understand what is occurring where. Sometimes there is a detailed and precise information of an **event** (spatial phenomenon defined by its location and the characteristics that can be measured in both quantitative or qualitative scale) across the geographical area such that a continuous map can be built out of that information. While, most of the time, only discrete and unrepresentative information on the spatial event is collected, which doesnot allow to create a precise continuous surface of the event over entire geographic area of interest. In such scenario, we should focus on two things;

- 1) How can we best utilize the available discrete information to represent the event across entire area using the appropriate sampling strategies.
- 2) Once a good sampling strategy is employed, we should be able to estimate the occurrence of the event across the unsampled region in the entirety of study area using proper spatial estimation technique.

This chapter introduces some basic ideas on different types of **sampling** strategies in spatial context. In addition, this chapter also introduces various type of spatial statistics that are being used in **predicting** the occurrence of events in the unsampled locations.

## Learning Objectives

1. To be familiar with different types of spatial sampling
2. To understand the relationship between observations at spatial scale using autocorrelation and semivariogram

3. To be familiar with the methods of spatial interpolation to predict observations at unknown locations
4. To be familiar with the methods of spatial prediction using regression models

## Key Terms

Geostatistics, Sampling, Prediction, Spatial Autocorrelation, Estimation, Geostatistics, P-value, Alpha, Null Hypothesis, Semi-variance, Semi-variogram, Population, Sampling Design, Sampling Unit, Probability Sampling, Simple Random Sampling, Stratified Random Sampling, Homogeneous, Systematic Sampling, Cluster Sampling, Purposive/Adaptive Sampling, Representative Sampling, Unique Case Sampling, Sequential Sampling, Spatial Autocorrelation, Domain, Attributes, Moran's I, Geary's C, Polygon Data, Point Data, Rook's/Queen's Case, Lag, Lag Distance, Gaussian, Spherical, Exponential, Circular, Spatial Interpolation, Thiessen Polygon, Inverse Distance Weighting, Kriging, Auxiliary, Spatial Lag Model, Spatial Weight Matrix, Distance-based Approach, Nearest-neighbor, Response Variable, Predictor Variable, Spatial Error Model, Errorsarl, Spatialreg, Basal Area, Non-Probability Sampling, Linear Kriging, Simple Kriging, Ordinary Kriging, Universal Kriging, Co-Kriging, Non-Linear Kriging, Indicator Kriging, Probability Kriging, Disjunctive Kriging

### 10.1 Introduction

Like any other data, spatial data will only make true sense if it can represent the entire population of interest and provide some flexibility in level of spatial analysis that can be performed from it. A representative sample will allow us to make unbiased and more accurate inferences of certain attribute of the population. While, spatial analysis will allow to understand where and what is occurring in the area of our interest. Two kind of limitation are inherent in a spatial data:

1. There is a high chance that the information regarding an attribute may be similar at nearby locations that are sampled due to **spatial autocorrelation**.
2. On the other hand, a spatial data carry the unsampled locations within a study area.

To overcome this limitation, one should come up with a proper **sampling** strategies to ensure that the given sample is representative of a population and lacks any redundant information. While the need to account for additional variables about a location may be intimidating, many spatial statistic analyses are out there to help with basic **prediction** and **estimation** of the information in an unobserved location. For example, interpolation and spatial regression can

help us predict and estimate the value of a variable in an unsampled location. Similarly, **spatial autocorrelation** measures the degree of similarity between samples at different sampled locations.

## Recall This

### 10.2 Geostatistics

Branch of statistics used to analyze and predict the values associated with spatial or spatiotemporal phenomena often defined by the locations over a study area.

### 10.3 Spatial Autocorrelation

The measures the degree of similarity between samples at different sampled locations.

### 10.4 Sampling

The process of selecting a part of a population.

### 10.5 Prediction

Using the existing sampled data to compute the value of the variable in a unsampled location.

### 10.6 Estimation

Using the existing data to find the best value of the coefficient while establishing the association between different variables

### 10.7 Classical vs. Geostatistical Inferences

In classical statistics, the variance is assumed to be totally random when samples are drawn from a defined sampling space [Jacquez, 1999] and are assumed to come from one distribution [Steel and Torrie, 1980]. The inference about the population is based on the comparison of a test statistic calculated for a sample to the distribution of the statistic under the null hypothesis for the reference distribution [Jacquez, 1999]. The significance level of a test statistic from the sample is compared with the critical value of test statistics obtained by repeatedly reordering the data at random proportion [Jacquez, 1999]. Interpretation of this significance level on the classical model is done using a **p-value**

which, if less than or equal to the a level of cutoff, also know as **alpha** (usually 5% or 0.05) the **null hypothesis of ‘no difference’** is rejected [Jacquez, 1999].

While in geostatistics, the variance is assumed to be partly random and each point in the field represents a sample from some distribution [Jacquez, 1999]. However, the distribution at any one point may differ completely from that at all other points in its shape, mean, and variance [Jacquez, 1999]. The distribution of differences in sample values separated by a specified distance is assumed to be the same over the entire field [Jacquez, 1999]. In geostatistics, if the sample values are similar (also know as **spatial autocorrelation**) given that they are in proximity to each other, the random variance of the distribution of differences in sample values is relatively small [Jacquez, 1999]. If the sample values doesnot show spatial autocorrelation, the variance is larger. In geostatistics, the **semi-variance**, i.e., half of the variance is used to measure the similarity between points at a given distance apart [Jacquez, 1999]. Using the semivarince, a graph of semivariance versus distance is constructed, know as **semi-variogram** [Jacquez, 1999]. Thus, beside conducting a hypothesis test , in geostatistics, there are two additional stages: 1) semivariogram construction and 2) incorporating semi-variogram to estimate the values at unsampled locations using various spatial methods. We will get into details about semivariogram construction and spatial methods in the later sections of this chapter.

## 10.8 Sampling

**Sampling** can be defined as the process of selecting some part of a population (also known as **sample**) in order to make an inference, and estimate some parameters about the whole population[Thompson, 2012]. For example, to estimate the amount of biomass of trees in Malcom Knapp Forest, BC, scientist collects the data on tree size, and height from randomly distributed 100 small plots across the forest. Based on some equations and using the tree data on height and size from 100 plots, biomass of entire Malcom Knapp Forest can be estimated. Similarly, to estimate the amount of recoverable oil in a region, a few (highly expensive) sample holes are drilled (example adapted from Thompson [2012]). The situation is similar in a national opinion survey, in which only a sample of the people in the population is contacted, and the opinions in the sample are used to estimate the proportions with the various opinions in the whole population (example adapted from Thompson [2012], pp.1).

**Sampling** should not be confused with **observational** study. In an observational study, one has no control over the sample while in sampling one can deliberately select samples which prevent research bias[Thompson, 2012]. Sampling accounts on how data can be collected out of every possibilities under the control of investigator [Thompson, 2012].

Broadly, sampling can be categorized into two groups [Teddlie and Yu, 2007]:

1. Probability sampling

## 2. Non-probability sampling

Before getting into the details about different types of sampling. We will make ourself familiar with some sampling key terms and their definitions.

### Recall This

## 10.9 Population

Any large spatially defined entity of plots, people, trees, animals etc., from which samples are drawn and measurement of certain characteristics is conducted.

## 10.10 Sampling Design

The procedure by which the sample of units is selected from the population is called the sampling design.

## 10.11 Sampling Unit

The smallest entity within a population from which the information about population is drawn is known as sampling unit. For example, in a survey of potential internet user over entire BC , sampling unit can be the certain number of household in each city across BC.

## 10.12 Probability Sampling

Probability sampling techniques are mostly used in studies that use extensive amount of quantitative analysis [Tashakkori and Teddlie, 2010]. It involves selecting a large number of units from a population where the probability of inclusion for every member of the population is determinable [Tashakkori and Teddlie, 2010].

## 10.13 Simple Random Sampling

In simple random sampling, each sampling unit within a given population has equal probability of being selected in a sample [Thompson, 2012].For example, suppose we would like to measure the tree heights of all the trees from a simple random sample of 60 plots with their spatial locations (given by plots center co-ordinates) from a forest divided into 625 spatially defined plots as given in **Figure 10.1**. Notice, there is no distinctive pattern on how plots are being selected for the measurement of tree heights, this justify the **random part** of the simple random sampling.

As an investigator, when we make a sequence of selections from a population, at each step, new and distinct set of sampling units are being selected in the sample, each having equal probability of being selected at each step.

For example, when we take another sample of 60 plots, we can see that different **sampling units (plots)** are being selected from what we obtained, this represent the **equal probability** of each sampling unit being selected.

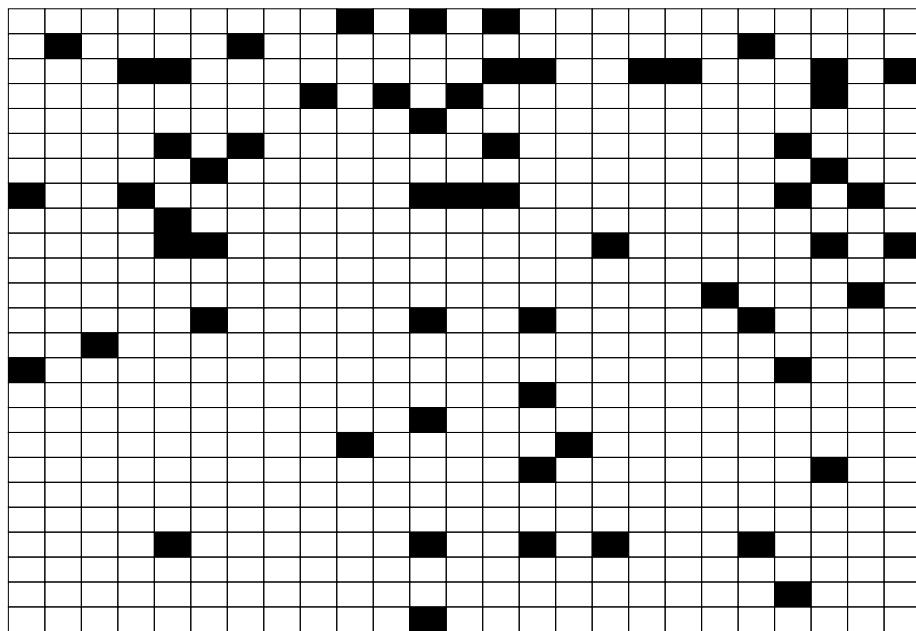


Figure 10.1: Simple random sample of 60 units from a population of 625 units.  
Nepal, CC-BY-SA-4.0.

Simple random sample of 60 units from a population of 625 units.

Another simple random sample of 60 units.

## 10.14 Stratified Random Sampling

When a population under study is not **homogeneous** (similar in biological characteristics) across the entire study area and consists of some sort of gradient, stratified random sampling method is used [Thompson, 2012]. The principle of stratification is to partition the population in such a way that the units within a stratum are as similar as possible [Teddlie and Yu, 2007]. Random samples from each strata are drawn to ensure adequate sampling of all groups [Teddlie and Yu, 2007]. Even though one stratum may differ markedly from another, a stratified sample with the desired number of units from each stratum in the

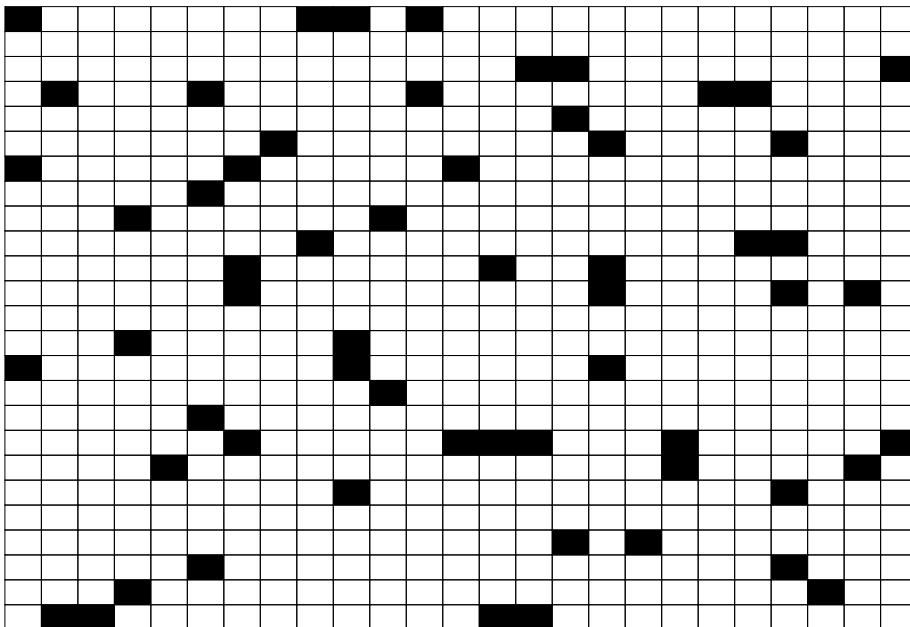


Figure 10.2: Another simple random sample of 60 units. Nepal, CC-BY-SA-4.0.

population will tend to be “representative” of the population as a whole [Howell et al., 2020].

For example, a forest under study is divided (**stratified**) into similar regions Figure 10.2 defined by elevation, soil moisture, and soil nutrient gradient and random samples are taken within each strata. The stratification of a study region despite of its size can help to spread the sample over the entire study area.

Stratified random sample within unequal strata within a study area.

Stratified random sample from equal strata within a study area.

## 10.15 Systematic Sampling

A systematic sample uses a fixed grid or array to assign plots in a regular pattern **Figure 10.4** [McRoberts et al., 2014]. The advantage of systematic sampling is that it maximizes the average distance between the plots and therefore minimizes spatial correlation among observations and increases statistical efficiency [McRoberts et al., 2014]. In addition, a systematic sample, which is clearly seen to be representative in some sense, can be very convincing to decision-makers who lack experience with sampling [McRoberts et al., 2014]. Raster grids such as digital elevation models (DEM) are some examples of systematic sample.

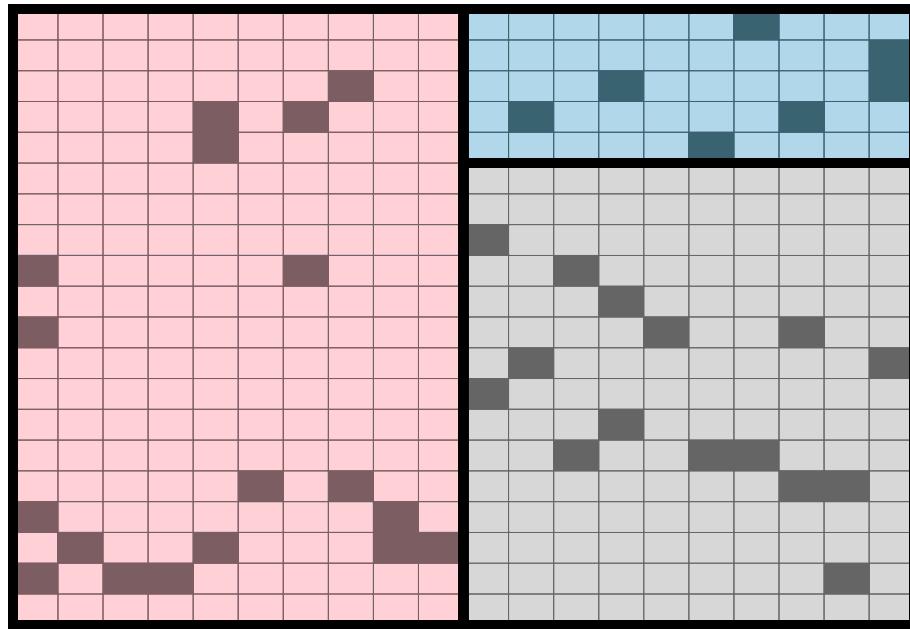
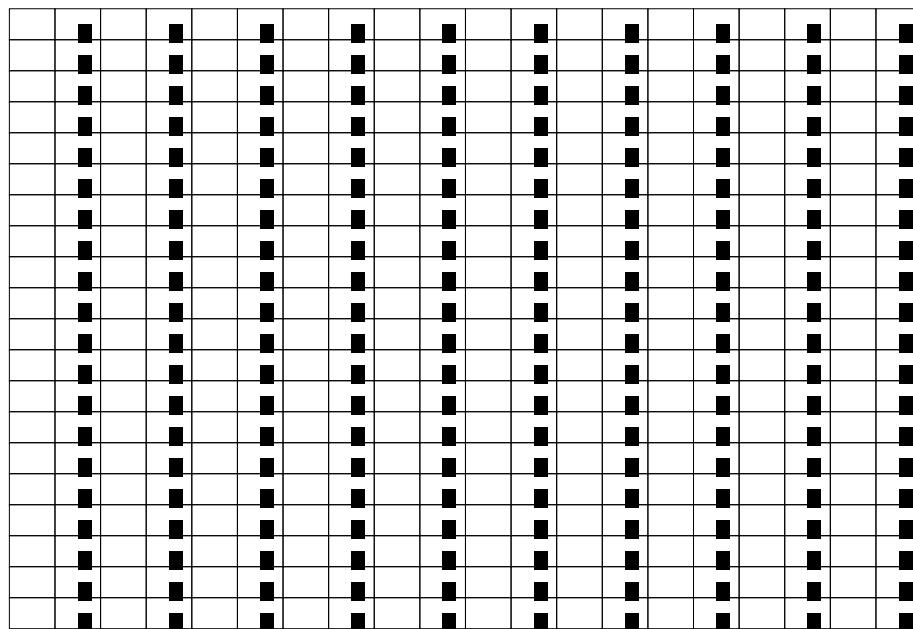


Figure 10.3: Stratified random sample within unequal strata within a study area. Nepal, CC-BY-SA-4.0.

Sample every second observation in the row



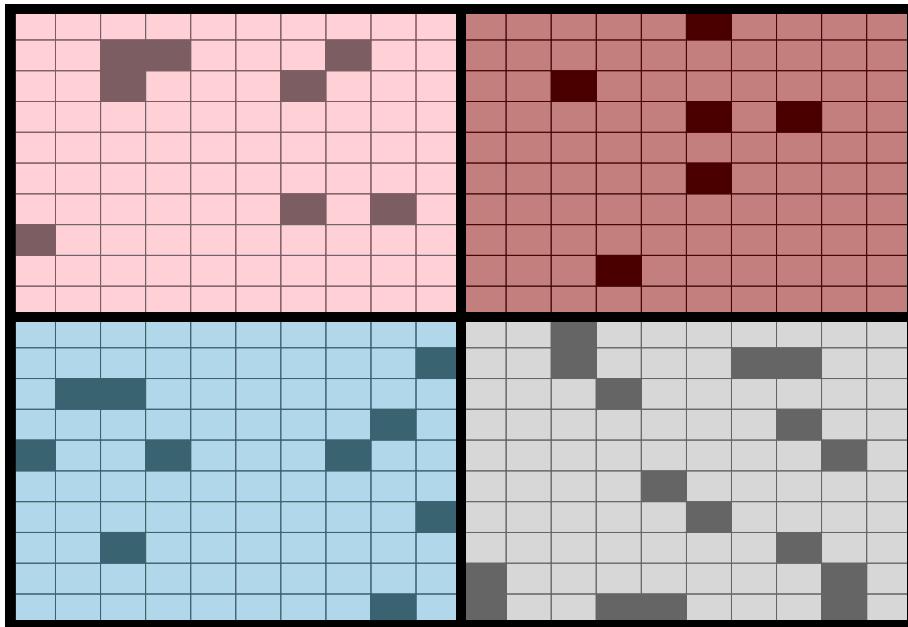


Figure 10.4: Stratified random sample from equal strata within a study area.  
Nepal, CC-BY-SA-4.0.

Sample all the observation in every second column

## 10.16 Cluster Sampling

In cluster sampling, rather than sampling individual units, which might be geographically spread over great distances, we can sample groups (clusters) of plots that occur naturally in the study area [Teddlie and Yu, 2007]. Cluster sampling is employed when we want to be more efficient in terms of the use of time and money to generate a more efficient probability sample [Teddlie and Yu, 2007].

Cluster of plots selected from entire study area

## 10.17 Non-probability Sampling

Non-probability sampling is generally used in qualitative studies. They are also known as **purposive or adaptive sampling**, and defined as selecting units (e.g., individuals, groups of individuals, institutions) based on specific purposes associated with answering some research questions. **Purposive or Adaptive** sampling can be classified into three broad categories [Teddlie and Yu, 2007]:

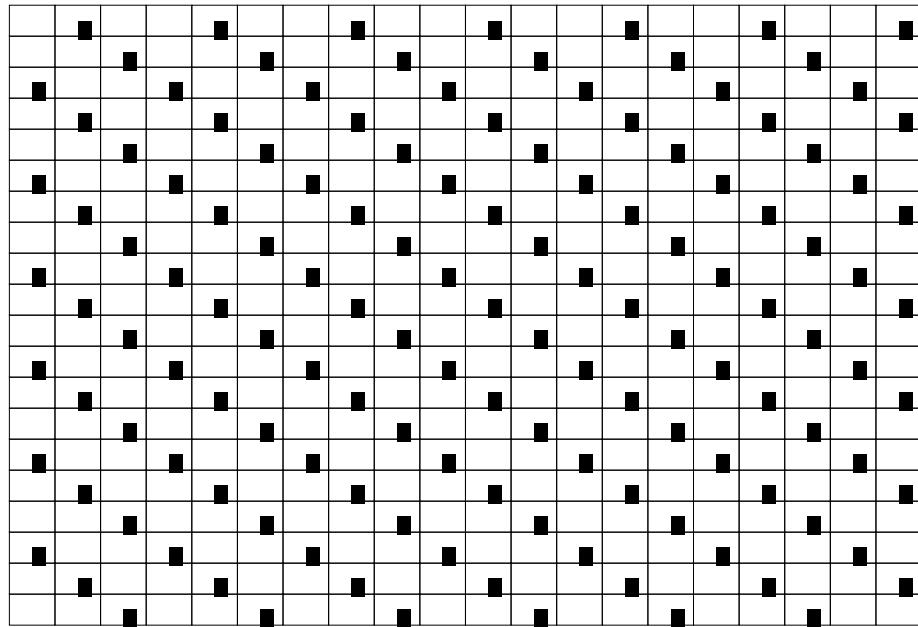


Figure 10.5: Sample every second observation in the row. Nepal, CC-BY-SA-4.0.

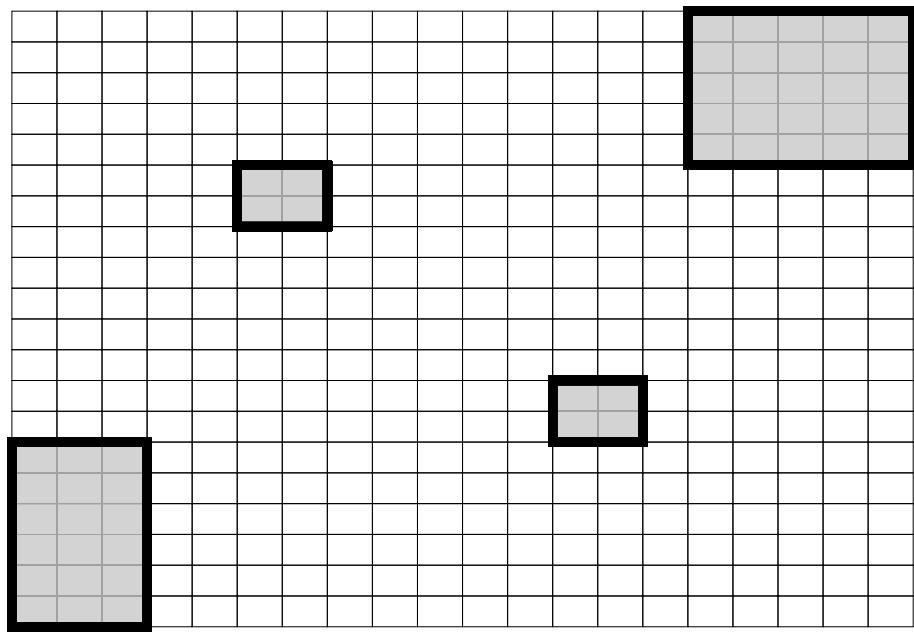


Figure 10.6: Cluster of plots selected from entire study area. Nepal, CC-BY-SA-4.0.

## 10.18 Representative Sampling

This type of sampling is used when we want to select samples that will represent broader groups as closely as possible [Teddlie and Yu, 2007]. One of the example of representative sampling is selecting 100 Douglas fir and 50 Spruce tree from study area within Malcom Knapp Forest, BC consisting of 500 Douglas fir and 300 Spruce trees for the measurement of tree height.

## 10.19 Unique Case Sampling

In this sampling, we want to focus on more specific case which is unique and rare in terms of one or more characteristics [Teddlie and Yu, 2007]. One of the example of unique case sampling could be understanding the genetic makeup of person who is not affected by Covid-19 virus.

## 10.20 Sequential Sampling

In this sampling method, we would pick up a single or group of cases in an interval of time, analyzes the results and then move on to the next group of cases and so on [Teddlie and Yu, 2007]. The goal of the research project is to generate some theory (or broadly defined themes) [Teddlie and Yu, 2007].

## 10.21 Spatial Autocorrelation

When an attribute or variable is mapped across a study area or domain, geologist ask a question on whether a variable is cluster, randomly distributed or dispersed [Carr and hsien Mao, 1993]. In some cases, the nature of a cluster is distinctive visually, while in others it is not apparent [Carr and hsien Mao, 1993]. Hence, to come up with a quantitative measure on variable is clustered or randomly distributed in the domain, **spatial autocorrelation** is used. The concept of autocorrelation comes from Tobler's first law of geography which states, "things that are closer in distance are related" [Tobler, 1970]. **Spatial autocorrelation** can be defined as the relationship between a variable of interest with itself when measured at different location [Cliff, 1973]. In other words, if a variable is measured at different locations which are at proximity, the value of variable is almost same. There could be both positive and negative spatial autocorrelation. Consider the following example; **Figure 10.6** shows the clustering pattern in the given square boxes (can be variable of interest) representing the positive spatial autocorrelation (left) and a complete checkerboard (right) distribution of square boxes (variable of interest) indicating a negative spatial autocorrelation.

## Recall This

### 10.22 Domain

The study area from where spatial sample is taken.

### 10.23 Attributes

The information attached to the study objects that are spatially distributed in a Domain. Often termed as variable of interest.

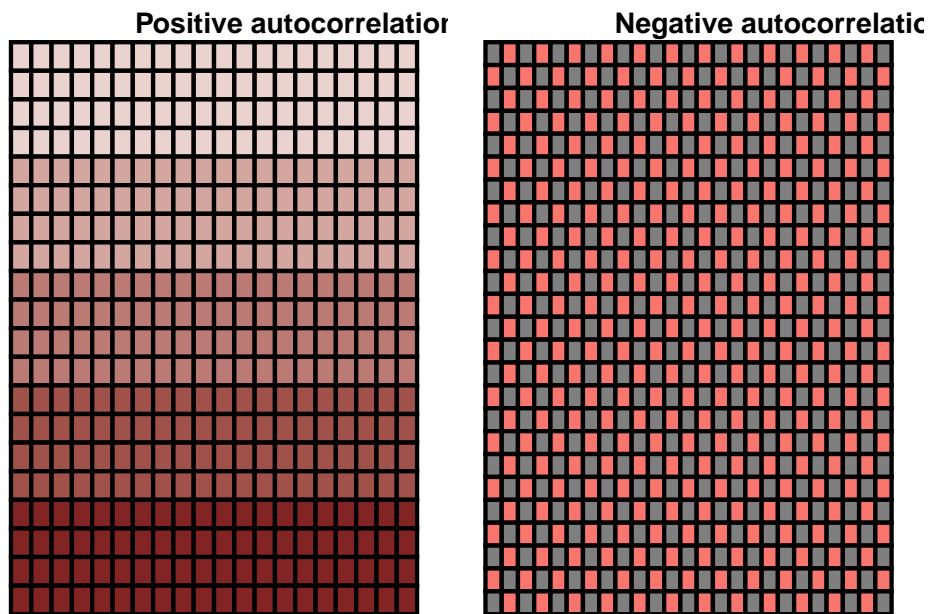


Figure 10.7: Example of a positive (left) and negative spatial autocorrelation (right) for a give domain. Nepal, CC-BY-SA-4.0.

Example of a positive (left) and negative spatial autocorrelation (right) for a give domain.

### 10.24 Moran's I

**Moran's I** [Moran, 1950], is a correlation coefficient that measures the degree of spatial autocorrelation in certain attributes of the data. It is based on the spatial covariance standardized by the variance of the data [Moran, 1950]. It works based on the neighborhood list created based on spatial weight matrix [Suryowati et al., 2018]. The value of Moran's I ranges between -1 to 1, where

1 indicates the perfect positive spatial autocorrelation, 0 indicates the random pattern, and -1 indicates the perfect negative autocorrelation [Moran, 1950]. Moran's I is calculated using the following formula [Moran, 1950]:

$$I = \frac{1}{s^2} \frac{\sum_i \sum_j (y_i - \bar{y})(y_j - \bar{y})}{\sum_i \sum_j w_{ij}}$$

Where,

$I$  = the Moran I statistics

,

$y_i$  = variable measure at location i

$y_j$  = variable measure at location j

$S^2$  = the variance

$w_{ij}$  = the spatial weight matrix

## 10.25 Case Study: Title of Case Study Here

You see textual case study content here

For this case study, we will use ground plot data from Change Monitoring Inventory (CMI) program [for details: Province of BC, 2018] for Williams Lake and 100-miles House timber supply area (TSA) in the province of British Columbia, Canada. William Lake TSA and 100-miles House TSA are divided into 18 and 8 blocks respectively **Figure 10.7**. There is a total of 456 CMI plots used in this study **Figure 10.7**. The total basal area (m<sup>2</sup>/ha) is our variable of interest in this study. For each of the polygon in Williams lake and 100-miles house TSA, total basal area was calculated by taking the sum of the basal area of each CMI plots in each polygon. For this part of exercise, we want to understand if there is any spatial relationship (autocorrelation) between the total basal area measured in each polygon of TSA. We will quantitatively measure the presence or absence of **spatial autocorrelation** using **Moran's I** and **Geary's C**.

```
## OGR data source with driver: ESRI Shapefile
## Source: "/Users/evan/Documents/github/geomatics-textbook/data/10", layer: "Block_basa_area"
## with 26 features
## It has 21 fields
## Integer64 fields read as strings: Field1

## OGR data source with driver: ESRI Shapefile
## Source: "/Users/evan/Documents/github/geomatics-textbook/data/10", layer: "100_and_will"
## with 2 features
## It has 16 fields
```

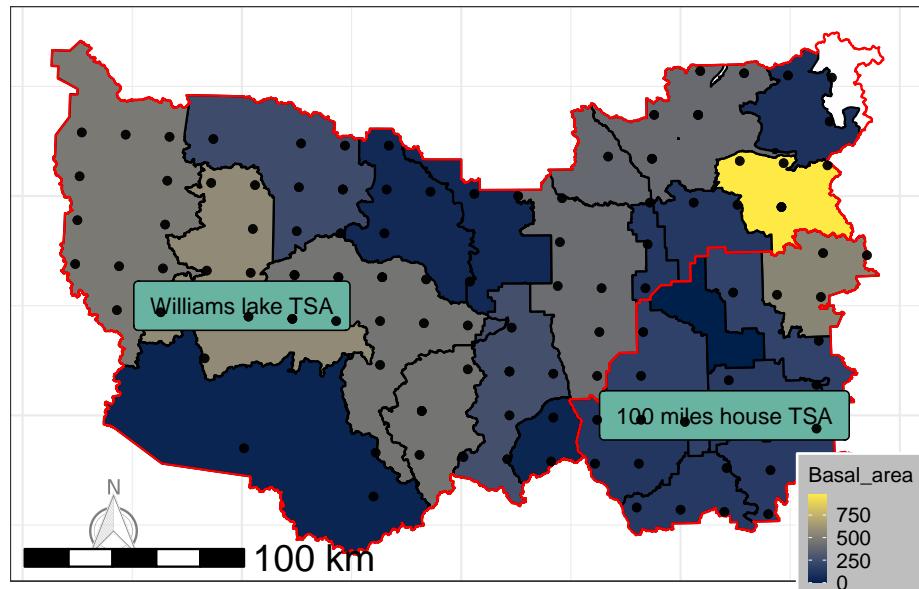


Figure 10.8: CMI plots location for williams lake TSA and 100 miles house TSA. Basal area has been calculated using the sum of the basal area for each plot (dot) over the given polygon. Nepal, CC-BY-SA-4.0.

CMI plots location for williams lake TSA and 100 miles house TSA. Basal area has been calculated using the sum of the basal area for each plot (dot) over the given polygon. :::

## Calculating Moran's I

We will calculate the Moran's I for the basal area variable pertaining to the polygon.

## Using Contiguity

### Define Neighborhood

The Moran's I statistic is the correlation coefficient for the relationship between a variable and its surrounding values. But before we go about computing this correlation, we need to come up with a way to define a neighborhood. There are two ways to define neighborhood namely; contiguity for spatial **polygon data** and distance-based approach for the spatial **point data** and polygon data both. For polygon data, contiguity based neighborhood selection can be adopted using two widely used method, respectively known as **Rook's case** or **Queen's case** **Figure 10.8**.

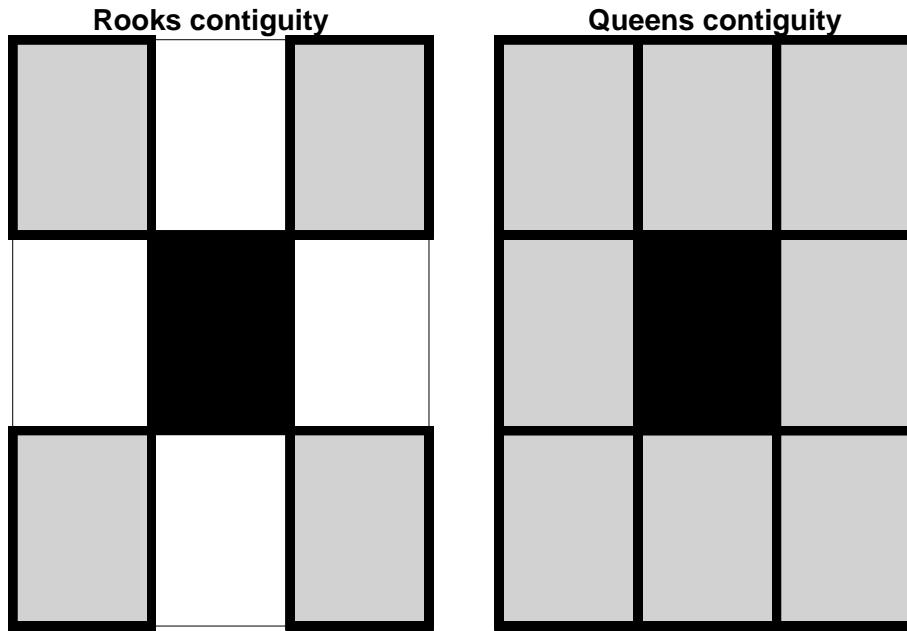


Figure 10.9: Rook's (left) and Queen's (right) case for searching the neighborhood (grey unit) for the darker unit in the center. Nepal, CC-BY-SA-4.0.

Rook's (left) and Queen's (right) case for searching the neighborhood (grey unit) for the darker unit in the center.

### Step 1: Build Neighborhood

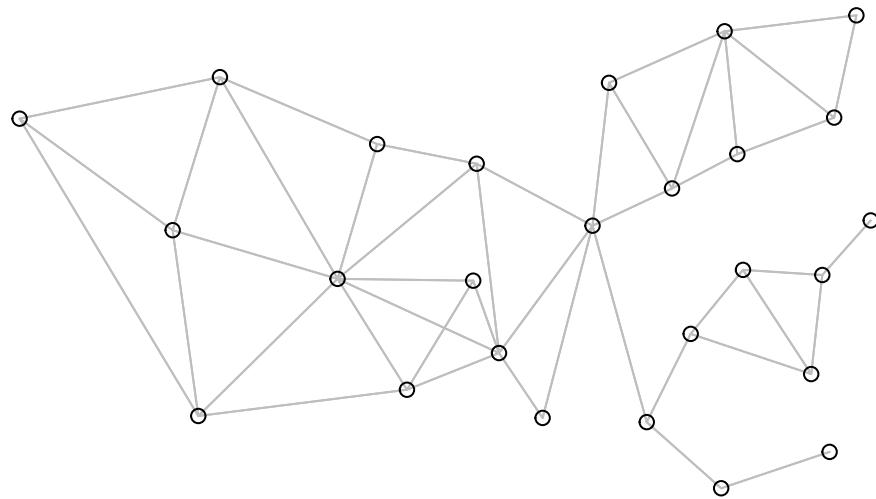
Since we are working with the polygons, we will use the queen's contiguity to build the neighborhood list using **poly2nb** function in **spdep** package in R and plot the linkage\*\*

```
#####
## Plot the data #####
## Convert the spatial data into data frame #####
filename_df <- tidy(filename)
# make sure the shapefile attribute table has an id column
filename$id <- rownames(filename@data)

# join the attribute table from the spatial object to the new data frame
filename_df <- left_join(filename_df,
                         filename@data,
                         by = "id")

#Searching neighborhood
w1 <- poly2nb(filename, row.names=filename$id, queen=T) ##### queens case
```

```
coords <- coordinates(filename)
plot(w1, coords, col="grey")
```



#### Step 2: Getting a Spatial Weight Matrix for Neighborhood List

```
ww <- nb2listw(w1, style='B')
```

#### Step 3: Calculate Moran's Correlation Coefficient Using the Spatial Weight Matrix for Neighbors

```
## calculating Moran's I
moran(filename$Basal, ww, n=length(ww$neighbours), S0=Szero(ww))
```

```
## $I
## [1] -0.1112932
##
## $K
## [1] 4.854083
```

#### Step 4: Conduct the Significance Test for the Calculated Moran's I Value

```
moran.test(filename$Basal, ww)
```

```
##
## Moran I test under randomisation
##
## data: filename$Basal
## weights: ww
##
## Moran I statistic standard deviate = -0.55645, p-value = 0.711
## alternative hypothesis: greater
```

```
## sample estimates:
## Moran I statistic      Expectation      Variance
##          -0.11129321     -0.04000000    0.01641516
```

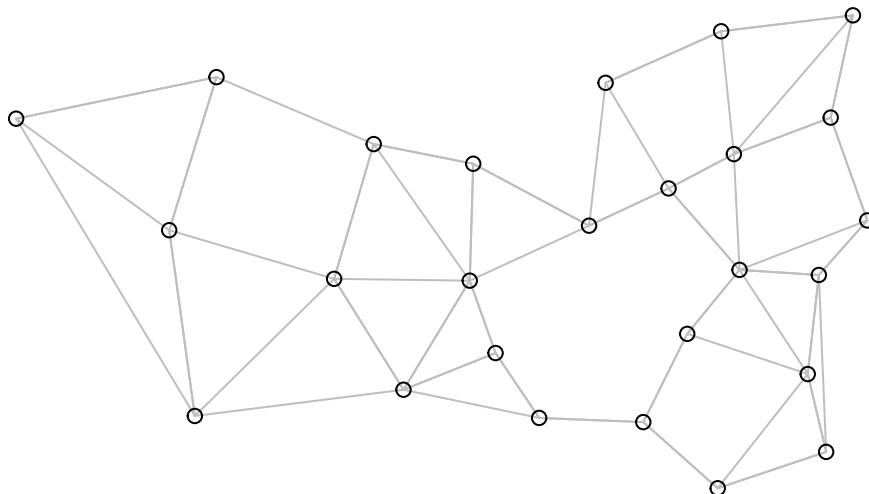
### Using the Knearest Neighborhood Imputation

Step 1: We will select 3 Nearest Neighbor Using Distance-based Approach

The function knearneigh in 'spdep' package in R and build a spatial weight

```
# Searching neighborhood
col.knn <- knearneigh(coords, k=3)
w<-knn2nb(col.knn, row.names = filename$id)

coords <- coordinates(filename)
plot(w, coords, col="grey")
```



Step 2: Build a Spatial Weight Matrix for the Neighborhood List

```
#spatial weight
ww1 <- nb2listw(w, style='B')
#ww1
```

Step 3: Calculate the Moran's I Coefficient

```
## Calculating Moran's I
moran(filename$Basal, ww1, n=length(ww1$neighbours), S0=Szero(ww1))

## $I
## [1] 0.01767042
##
## $K
## [1] 4.854083
```

**Step 4: Significance Test for Calculated Moran's I**

```

moran.test(filename$Basal, ww1)

##
## Moran I test under randomisation
##
## data: filename$Basal
## weights: ww1
##
## Moran I statistic standard deviate = 0.43976, p-value = 0.3301
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic      Expectation      Variance
##          0.01767042     -0.04000000     0.01719815

```

Note that the value of Moran's I changed based on how we calculated the neighborhood list using two different approach. The interpretation change here based on the way we created the neighborhood. With contiguity based neighbor, we found a negative value for I (-0.11), indicating a negative weak spatial autocorrelation. When we run the significance test we can see that the p-value < 0.05 indicating the autocorrelation is not significant. While using nearest neighbor we found that I (0.017) indicated a weak positive spatial autocorrelation. One reason for the difference is k-nearest neighbor uses polygon within a greater distance and can include more polygons as compared to contiguous neighbor which uses either "queens" or "rooks" contiguity [Suryowati et al., 2018].

## 10.26 Geary's C

Another more local measure of spatial autocorrelation unlike Moran's I is **Geary's C** [Geary, 1954]. While Moran's I is calculated by standardizing the spatial autocovariance by the variance of the data. Geary's c on the other hand uses the sum of the squared differences between pairs of data values as it is a measure of covariance [Geary, 1954]. However, both statistics depends on the spatial nature of data and are based on neighborhood. Both of these statistics depend on a spatial structure specified by a spatial weights matrix. The value of Geary's C ranges between 0 to some unknown positive value, where 0 indicates the spatial randomness, values less than 1 indicates the positive spatial autocorrelation, while value greater than 1 indicates negative spatial autocorrelation [Geary, 1954]. It is calculated using following formula:

$$C = \frac{(n - 1) \sum_i^n \sum_j^n w_{ij} (y_i - y_j)^2}{2 \sum_i^n \sum_j^n w_{ij} \sum_i (y_i - \bar{y})^2}$$

**Step 1 and 2: We will calculate all the neighborhood list exactly how we did for Moran's I and get our spatial weight matrix**

**Step 2:** In our final step, we will use geary funtion from “spdep” package to calculate the value of Geary's

*For Queens Case*

```
## Geary C
geary(filename$Basal, ww, n=length(ww$neighbours), n1=length(ww$neighbours)-1, S0=Szero(ww))

## $C
## [1] 0.9944114
##
## $K
## [1] 4.854083

## Significance test for Geary C
geary.test(filename$Basal, ww)

## 
## Geary C test under randomisation
##
## data: filename$Basal
## weights: ww
##
## Geary C statistic standard deviate = 0.02595, p-value = 0.4896
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic      Expectation      Variance
##          0.99441140      1.00000000      0.04638142
```

*For Nearest Neighbour Method*

```
## Geary C
geary(filename$Basal, ww1, n=length(ww1$neighbours), n1=length(ww1$neighbours)-1, S0=Szero(ww1))

## $C
## [1] 0.9416159
##
## $K
## [1] 4.854083

## Significance test for Geary C
geary.test(filename$Basal, ww1)

## 
## Geary C test under randomisation
##
## data: filename$Basal
## weights: ww1
##
## Geary C statistic standard deviate = 0.3791, p-value = 0.3523
```

```

## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic      Expectation      Variance
##          0.94161588     1.00000000    0.02371762

```

Note that the value of Geary's C indicated a positive spatial autocorrelation using both queens case and k-nearest neighbor. However, the spatial autocorrelation was not significant as given by p-value < 0.05. Both Moran's I and Geary's C are in agreement in terms of results.

## 10.27 Semivariogram Modeling

**Semivariogram** is a basic geostatistical tool for measuring spatial autocorrelation of a variable measured at different spatial location. A semivariogram is a measure of variance of a variable between two specified location separated by certain distance which is termed as **lag** distance [Isaaks and Srivastava, 1989]. For example, we can measure how a variable  $y$  changes in value between site  $i$  and  $j$  by calculating the difference  $y(i) - y(i + j)$ , where  $i+j=h$  is a lag distance. If the surface represented by the two points is continuous and  $j$  is a small distance, one expects the difference to be small [Isaaks and Srivastava, 1989]. With increasing  $j$ , the difference increases. Let's translate this intuitive statement into a formula:

$$\gamma_h = \frac{1}{2N} \sum_{i,j=1}^{N(h)} (y_i - y_j)^2$$

Where,

$\gamma_h$  = semivariance at a spatial lag  $h$

$i$  = measure spatial coordinate (latitude/UTM easting)

$j$  = measure spatial coordinate (longitude/UTM northing)

$y_i$  = measured value of variable of interest at the spatial location  $i$

$y_j$  = measured value of variable of interest at the spatial location  $j$

$N$  = number of sampled differences or lag

Like the familiar variance of basic statistics, it is a sum of squares divided by the number  $N$  of sampled differences. Unlike simple variance about a mean, the semivariogram measures difference between two samples. The 'semi' in semivariogram comes from the fact that the variance is divided by 2.

A **semivariogram** is a graph that consists of **semivariance** on the y-axis and a **lag distance** on the x-axis **Figure 10.10**. There are various components in a semivariogram that can be used to interpret the nature and structure of spatial autocorrelation. Various components of a semivariogram are:

Nugget (C): Nugget refers to an unaccounted autocorrelation due to a small lag distance than sampling distance or due to sampling errors **Figure 10.10**.

Range (R): The distance at which a variogram model first flattens out. This is the distance upto which y variables are spatially autocorrelated or aggregated **Figure 10.10**.

Sill (S): The value of semivariance that a variogram model attains at a given range is called sill **Figure 10.10**.

Partial sill: The sill minus nugget

$$S_i = S - R$$

Partial sill to total sill ratio: This is the structural variance explained by the fitted semivariogram model [Rossi et al., 1992]. Amount of variance that is spatially autocorrelated [Rossi et al., 1992].

$$\text{Ratio} = \frac{S_i}{S + R}$$

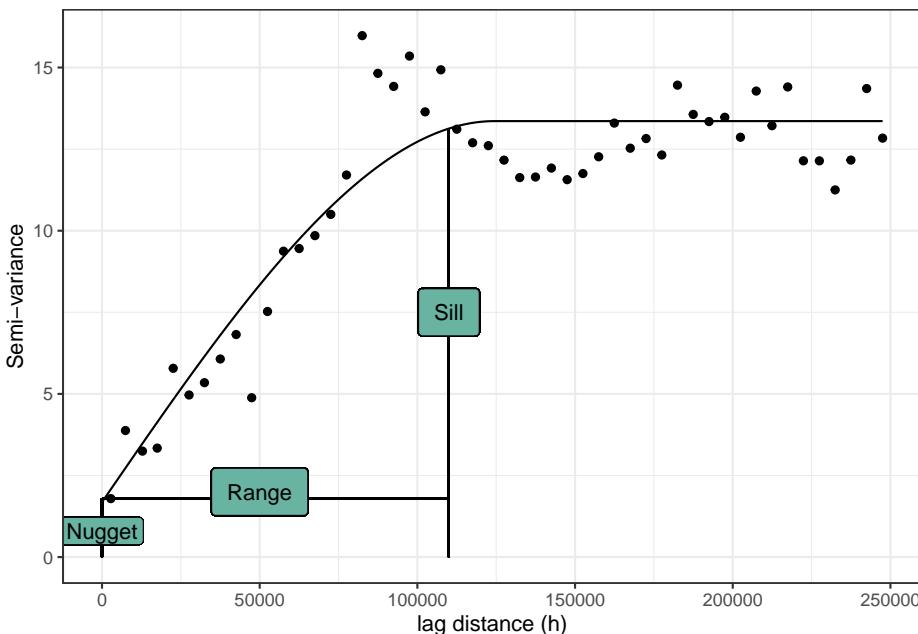


Figure 10.10: An example semivariogram with all the components using the ‘Fulmar’ data from ‘gstat’ package in R. Nepal, CC-BY-SA-4.0.

An example semivariogram with all the components using the “Fulmar” [see details Pebesma et al., 2005] data from “gstat” package in R.

## 10.28 Case Study: Title of Case Study Here

You see textual case study content here

For this case study, we will use the ground plot data from Young stand monitoring (YSM) program data [Province of BC, 2018] for Fort Saint Johns timber supply area (TSA) in the province of British Columbia, Canada. Fort Saint Johns is divided into 6 blocks respectively **Figure 10.11**. There is a total of 108 YSM plots used in this study **Figure 10.11**. The total basal area (m<sup>2</sup>/ha) is our variable of interest in this study. For each of the YSM plots, we will calculate the total basal area by adding the basal area for all trees within the plot. We will explore different type of semivariogram model with the same dataset and check which one will best fit the data.

::::

### 10.28.1 Gaussian

In actual application, Gaussian model usually suggest that at shorter lag distance the correlation is extremely high and it drops faster compared to any other variogram models.

```
fig_cap <- paste0("A semivariogram using the Gaussian model for the basal area (m2/ha)
## summarize the basal area at plot level
data1<- data%>%
  group_by(utm_easting,utm_northing) %>%
  summarise(Basal=sum(baha_L))
coordinates(data1)= ~ utm_easting+utm_northing

## Model formula
TheVariogram=variogram(Basal~1, data=data1)
## Initiating the parameters for the variogram , starting search window
TheVariogramModel <- vgm(psill=3000, model="Gau", nugget=100, range=20000)
## fitting a variogram model (Gaussian)
FittedModel <- fit.variogram(TheVariogram, model=TheVariogramModel)
preds = variogramLine(FittedModel, maxdist = max(TheVariogram$dist))

## Making some nice plots
g<-ggplot(TheVariogram,aes(x=dist,y=gamma))+geom_point()+
  geom_line(data = preds)+ theme_classic()+
  labs(x = "lag distance (h)", y = "Semi-variance")
g
```

A semivariogram using the Gaussian model for the basal area (m<sup>2</sup>/ha) for young stand monitoring plots.

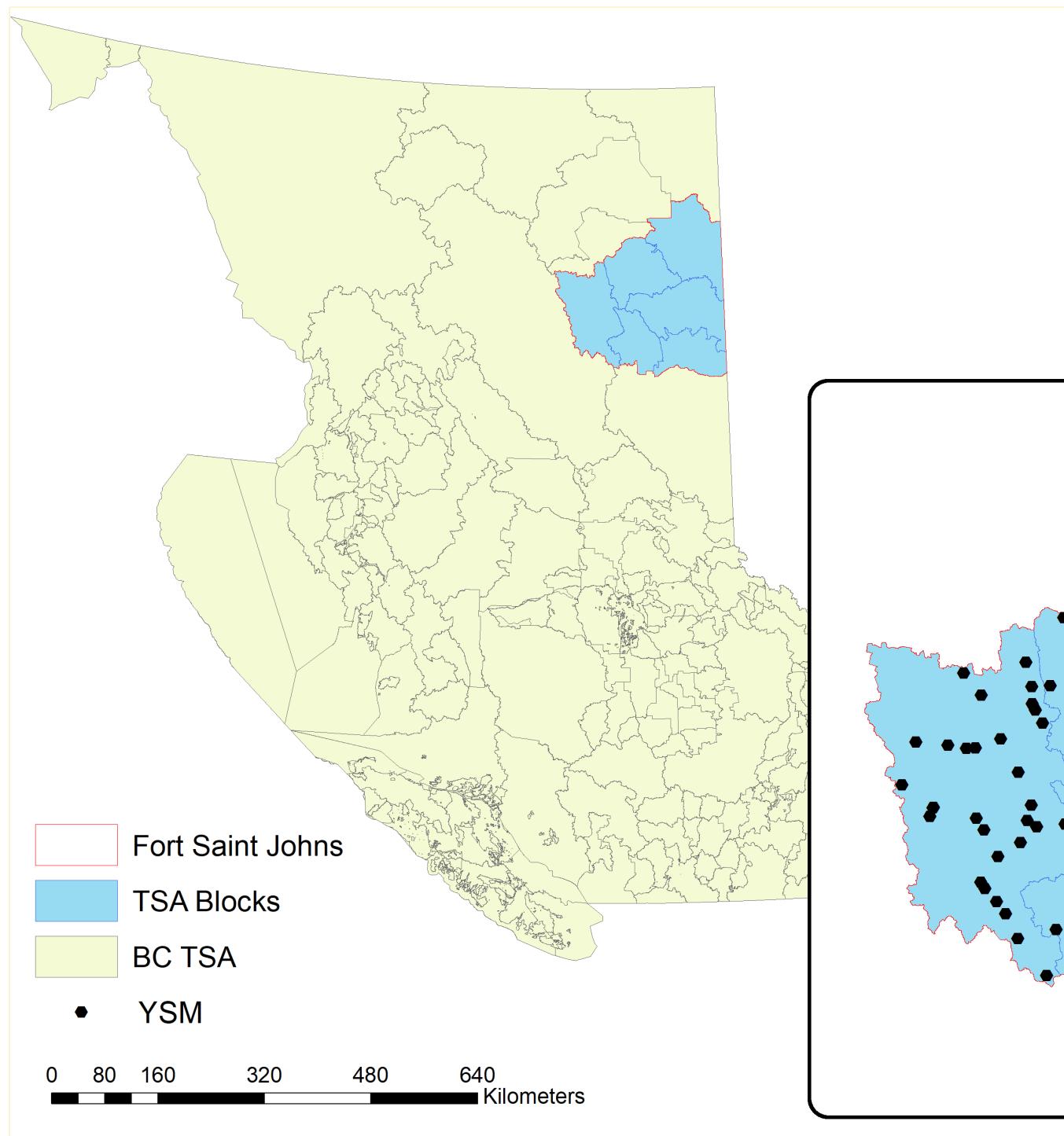


Figure 10.11: Location of Fort Saint Johns TSA and the young stand change monitoring plots. Nepal, CC-BY-SA-4.0.

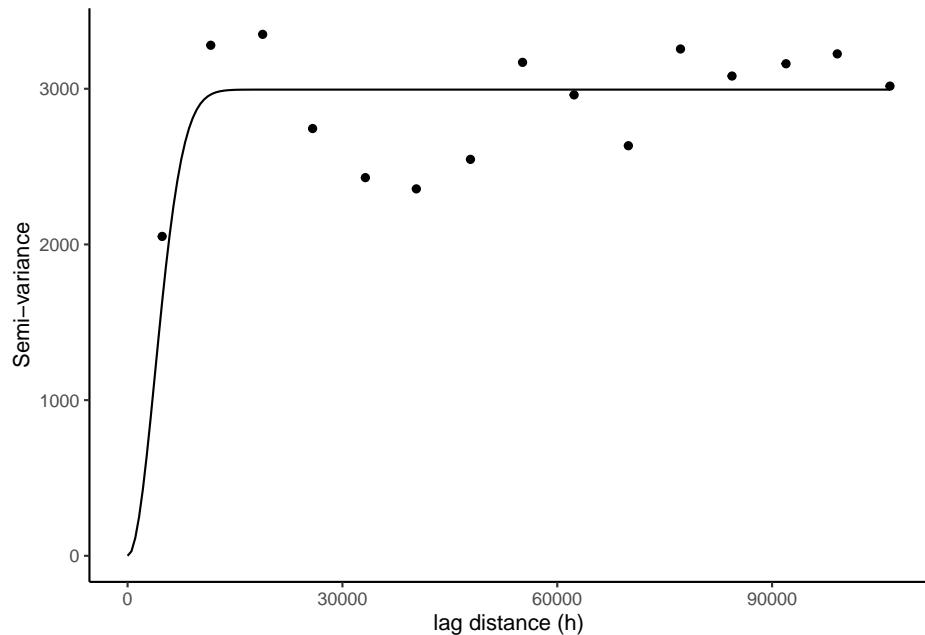


Figure 10.12: A semivariogram using the Gaussian model for the basal area (m<sup>2</sup>/ha) for young stand monitoring plots. Nepal, CC-BY-SA-4.0.

## 10.29 Spherical

This model shows a progressive decrease of spatial autocorrelation (equivalently, an increase of semivariance) until some distance, beyond which autocorrelation is zero. The spherical model is one of the most commonly used models.

```
fig_cap <- paste0("A semivariogram using the Spherical model for the basal area (m2/ha")
## summarize the basal area at plot level
data1<- data%>%
  group_by(utm_easting,utm_northing) %>%
  summarise(Basal=sum(baha_L))
coordinates(data1)= ~ utm_easting+utm_northing

## Model formula
TheVariogram=variogram(Basal~1, data=data1)
## Initiating parameters
TheVariogramModel <- vgm(psill=3000, model="Sph", nugget=60, range=20000)

## Fitting and predicting
FittedModel <- fit.variogram(TheVariogram, model=TheVariogramModel)
preds = variogramLine(FittedModel, maxdist = max(TheVariogram$dist))
```

```
## Some nice graph
g<-ggplot(TheVariogram,aes(x=dist,y=gamma))+geom_point()+
  geom_line(data = preds)+ theme_classic()+
  labs(x = "lag distance (h)", y = "Semi-variance")
g
```

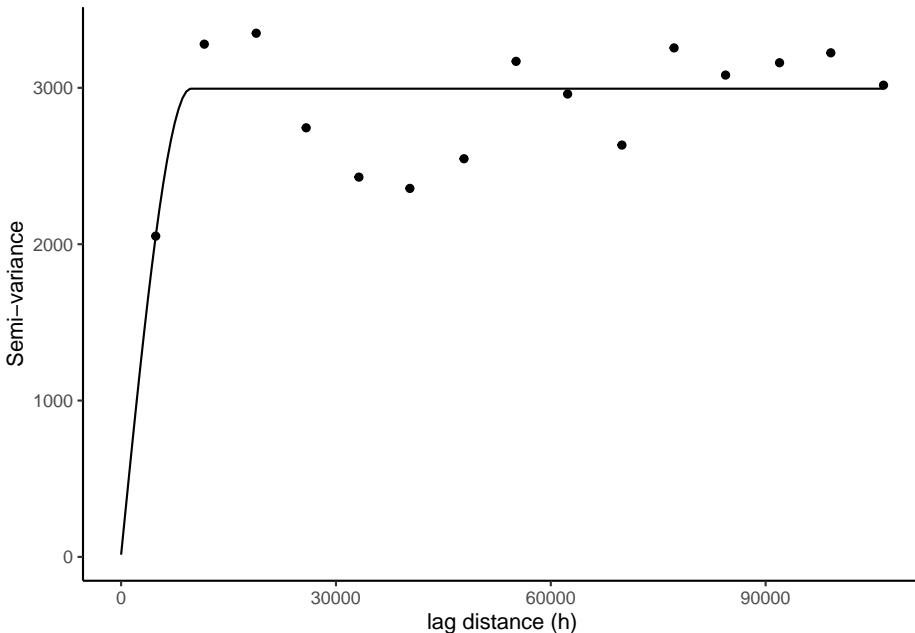


Figure 10.13: A semivariogram using the Spherical model for the basal area (m<sup>2</sup>/ha) for young stand monitoring plots. Nepal, CC-BY-SA-4.0.

A semivariogram using the Spherical model for the basal area (m<sup>2</sup>/ha) for young stand monitoring plots.

## 10.30 Exponential

Exponential model is used when spatial autocorrelation decreases exponentially with increasing distance. Here, the autocorrelation disappears completely only at an infinite distance.

```
fig_cap <- paste0("A semivariogram using the exponential model for the basal area (m2/ha) for for"
## summarize the basal area at plot level
data1<- data%>%
  group_by(utm_easting,utm_northing) %>%
  summarise(Basal=sum(baha_L))
coordinates(data1)= ~ utm_easting+utm_northing
```

```
## Model formula
TheVariogram=variogram(Basal~1, data=data1)
TheVariogramModel <- vgm(psill=3000, model="Exp", nugget=60, range=20000)
FittedModel <- fit.variogram(TheVariogram, model=TheVariogramModel)
preds = variogramLine(FittedModel, maxdist = max(TheVariogram$dist))
g<-ggplot(TheVariogram,aes(x=dist,y=gamma))+geom_point()+
  geom_line(data = preds)+ theme_classic()+
  labs(x = "lag distance (h)", y = "Semi-variance")
g
```

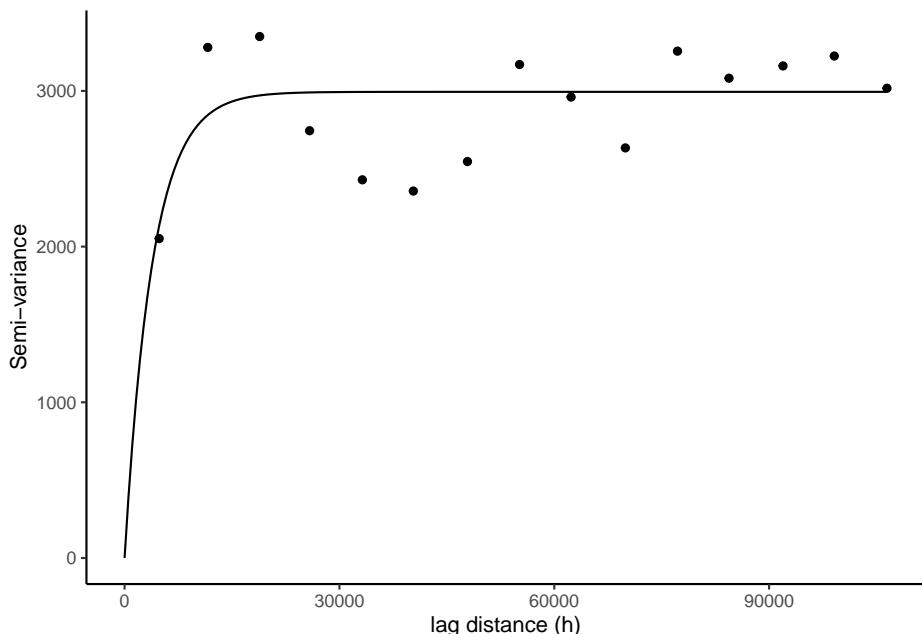


Figure 10.14: A semivariogram using the exponential model for the basal area (m<sup>2</sup>/ha) for young stand monitoring plots. Nepal, CC-BY-SA-4.0.

A semivariogram using the exponential model for the basal area (m<sup>2</sup>/ha) for young stand monitoring plots.

### 10.31 Circular

A circular variogram means that there is no preferred orientation in the data. We are only interested in the values of the variable of interest without considering the spatial orientation of the data.

```

fig_cap <- paste0("A semivariogram using the circular model for the basal area (m2/ha) for for yo
## summarize the basal area at plot level
data1<- data%>%
  group_by(utm_easting,utm_northing) %>%
  summarise(Basal=sum(baha_L))
coordinates(data1)= ~ utm_easting+utm_northing

## Model formula
TheVariogram=variogram(Basal~1, data=data1)
TheVariogramModel <- vgm(psill=3000, model="Cir", nugget=60, range=20000)
FittedModel <- fit.variogram(TheVariogram, model=TheVariogramModel)
preds = variogramLine(FittedModel, maxdist = max(TheVariogram$dist))
g<-ggplot(TheVariogram,aes(x=dist,y=gamma))+geom_point()+
  geom_line(data = preds)+ theme_classic()+
  labs(x = "lag distance (h)", y = "Semi-variance")
g

```

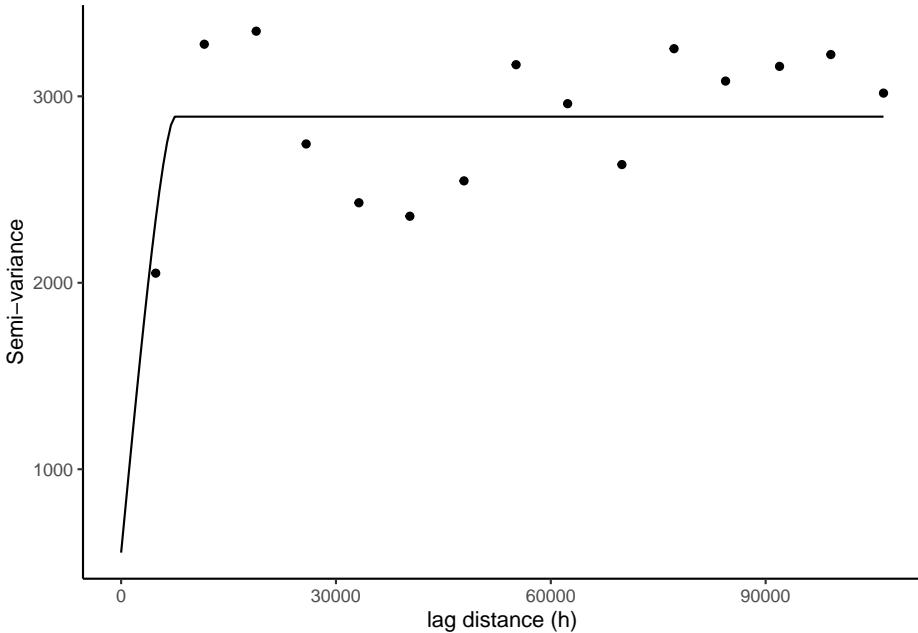


Figure 10.15: A semivariogram using the circular model for the basal area (m<sup>2</sup>/ha) for for young stand monitoring plots. Nepal, CC-BY-SA-4.0.

A semivariogram using the circular model for the basal area (m<sup>2</sup>/ha) for for young stand monitoring plots.

Just looking at the variograms, it appears that all of the four models fit our data well and indicates there is a strong correlation in basal area per hectare

of live trees between the plots. However, we will use all the components of semivariogram models to pick our best fitting variogram.

**Table 10.1** Summary of various component of variogram for four different models.

Model	Range	Nugget	Sill	Partial_sill	Sill_to_Sill
Circular	7433.68	552.71	2338.27	1785.56	0.76
Gaussian	4446.62	0.00	2995.05	4446.62	1.00
Spherical	9729.37	14.13	2980.05	2966.37	0.99
Exponential	3871.76	0.00	2994.05	2994.05	1.00

We can see that the partial sill to total sill ratio is highest for Gaussian and Exponential variogram. This indicated a highest total amount of semi-variance that is spatially correlated. Similarly, we can see that both models are indicating that we are observing a spatial autocorrelation in basal area at a very shorter range. Since with, exponential variogram autocorrelation only disappear at a infinite distance in reality , it is better to pick Gaussian model in most of the similar cases like we have in this study.

## 10.32 Spatial Interpolation

**Spatial interpolation** can be defined as the process of predicting the given **variable of interest** at an unmeasured location given we have the sample in the proximity of the unknown location. Spatial interpolation methods can be categorized into two broad groups:

1. Methods without using semivariogram
2. Methods using semivariogram

We will discuss both method with a case study in detail.

## 10.33 Case Study: Title of Case Study Here

You see textual case study content here

For this case study, we will use ground plot data from Young stand monitoring (YSM) program data [Province of BC, 2018] for Fort Saint Johns timber supply area (TSA) in the province of British Columbia, Canada.Fort Saint Johns is divided into 6 blocks respectively **Figure 10.11**. There is a total of 108 YSM plots used in this study **Figure 10.11**. The total basal area (m<sup>2</sup>/ha) is our variable of interest in this study. For each of the YSM plot the total basal area was calculated by adding the basal area for all trees within the plot. We will use this dataset to explore different interpolation technique to find the variable of interest (basal area) within the unsampled locations.

::::

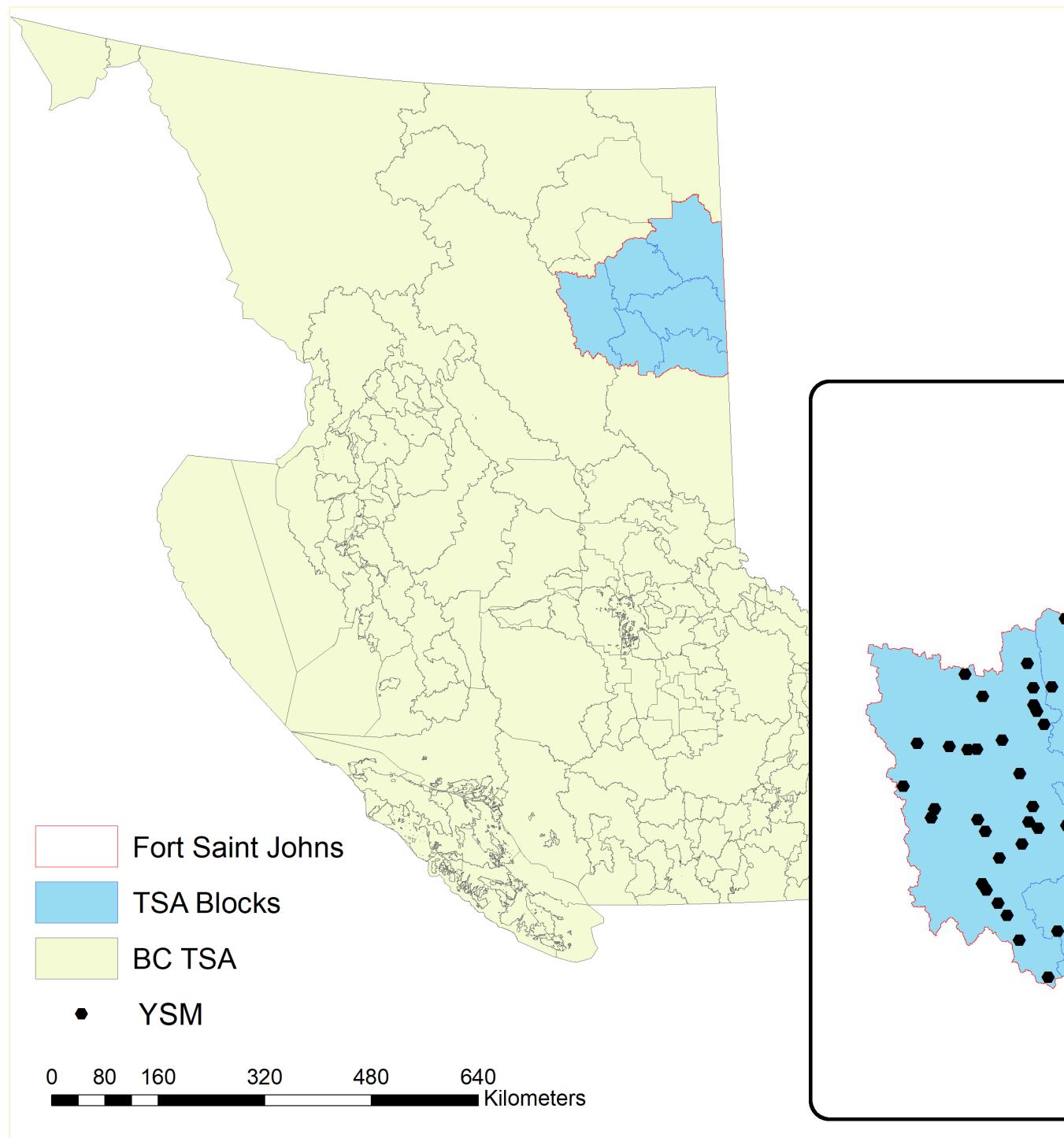


Figure 10.16: Location of Fort Saint Johns TSA and the young stand change monitoring plots. Nepal, CC-BY-SA-4.0.

### 10.34 Methods Without Using Semi-variogram

```
filename <-read_sf(dsn="data/10",layer="Fort_St_Jh")
plot1<- read.csv("data/10/FJS_plots.csv",header=T)
```

### 10.35 Nearest Neighbor

Nearest neighbor interpolation approach uses the value of variable of interest from the nearest sampled location and assign the value to the unsampled location of interest [Titus and Geroge, 2013]. It is very simple method and is most widely used for image processing in remote sensing research [Titus and Geroge, 2013].

**Step 1:** Match the projection of plot data with the study area boundary

```
# spatstat Used for the dirichlet tessellation function
# maptools Used for conversion from SPDF to ppp
# raster Used to clip out thiessen polygons
spdf <- as_Spatial(filename)
dsp <- SpatialPoints(plot1[,14:15], proj4string=CRS("+proj=utm +zone=10 +ellps=GRS80 +towgs84=0,0,0"))
dsp <- SpatialPointsDataFrame(dsp, plot1)
#####
TA <- CRS("+proj=utm +zone=10 +ellps=GRS80 +datum=NAD83")
library(rgdal)
dta <- spTransform(dsp, TA)
cata <- spTransform(spdf, TA)
```

**Step 2:** Create polygons throughout the study area where interpolation is to be done and rasterize the polygons

```
fig_cap <- paste0("An intermediate step in creating polygon and rasterizing it over the entire Fort Saint Johns TSA")
v <- voronoi(dta)
plot(v)

vca <- intersect(v, cata)
spplot(vca, 'baha_L', col.regions=rev(get_col_regions()))

#####
r <- raster(cata, res=100)
vr <- rasterize(vca, r, 'baha_L')
```

An intermediate step in creating polygon and rasterizing it over the entire Fort Saint Johns TSA.

**Step 3:** Nearest neighbor with five unsampled points to be interpolated at a time and plot the results

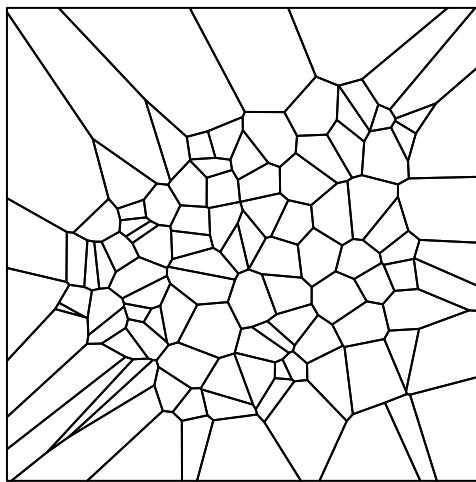


Figure 10.17: An intermediate step in creating polygon and rasterizing it over the entire Fort Saint Johns TSA. Nepal, CC-BY-SA-4.0.

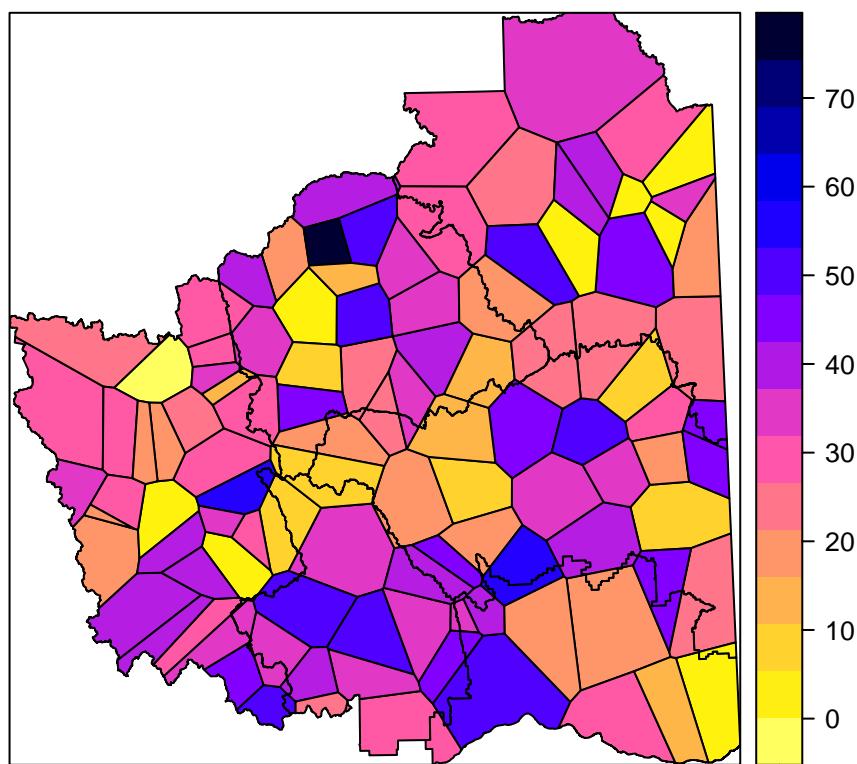


Figure 10.18: An intermediate step in creating polygon and rasterizing it over the entire Fort Saint Johns TSA. Nepal, CC-BY-SA-4.0.

```

fig_cap <- paste0("Predicted basal area over the entire Fort Saint Johns TSA using five
## gstat package to create semivariogram model, kriging an dinterpolation
gs <- gstat(formula=baha_L~1, locations=dta, nmax=5, set=list(idp = 0))
nn <- interpolate(r, gs)

## [inverse distance weighted interpolation]
nnmsk <- mask(nn, vr)
tm_shape(nnmsk) +
  tm_raster(n=8,palette = "RdBu", auto.palette.mapping = FALSE,title="Predicted basal a
  tm_legend(legend.outside=FALSE)

```

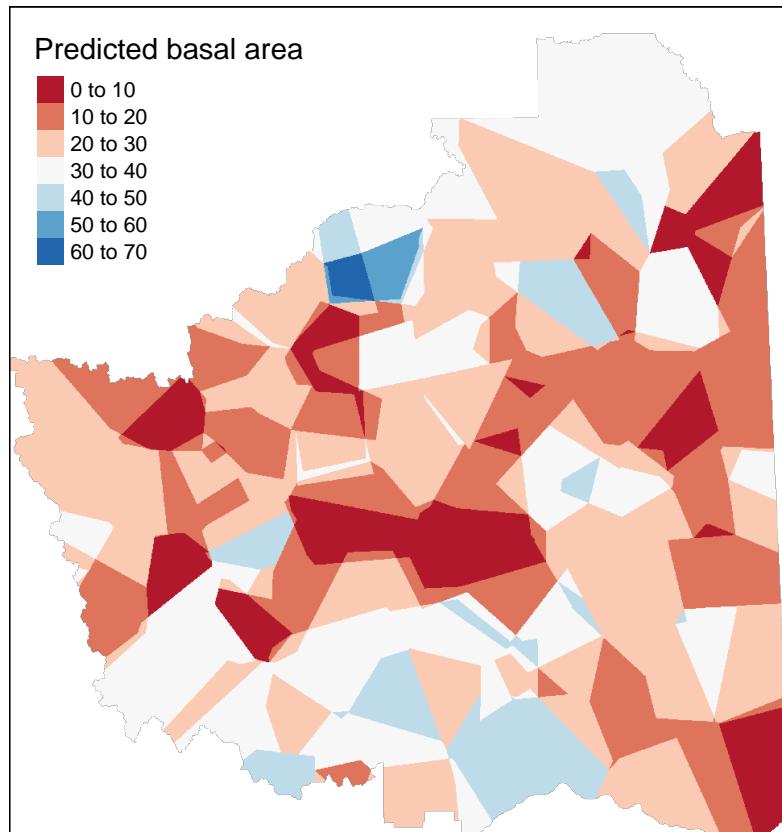


Figure 10.19: Predicted basal area over the entire Fort Saint Johns TSA using five nearest neighbor. Nepal, CC-BY-SA-4.0.

Predicted basal area over the entire Fort Saint Johns TSA using five nearest neighbor.

#### Step 4: Leaflet map for some interactions

Predicted basal area over the entire Fort Saint Johns TSA using five nearest neighbor projected over the province of British Columbia. Nepal, CC-BY-SA-4.0.

Predicted basal area over the entire Fort Saint Johns TSA using five nearest neighbor projected over the province of British Columbia.

## 10.36 Thiessen Polygon

In this method, the domain is determined into the area/polygons of regions containing one sampling point from the original data [Coulston and Reams]. The thiessen polygons are assigned with the same values of the variable of interest as the point sampled [Yamada, 2016].

### Step 1: Match the projection of the Shape file and the plot data

```
# project the plot data based to UTM zone 10 and NAD83
dsp <- SpatialPoints(plot1[,14:15], proj4string=CRS("+proj=utm +zone=10 +ellps=GRS80 +datum=NAD83")
# convert the data into spatial object
dsp <- SpatialPointsDataFrame(dsp, plot1)
## change the projection of both shape file and plot data
TA <- CRS("+proj=utm +zone=10 +ellps=GRS80 +datum=NAD83")
library(rgdal)
dta <- spTransform(dsp, TA)
cata <- spTransform(spdf, TA)
```

### Step 2: Create the thiessen polygon around the sample points for entire TSA using

```
#"dirichlet" function from "spatstat" package
# Create a tessellated surface
th <- as(dirichlet(as.ppp(dta)), "SpatialPolygons")

# The dirichlet function does not carry over projection information
# requiring that this information be added manually to the thiessen polygons
proj4string(th) <- proj4string(dta)
```

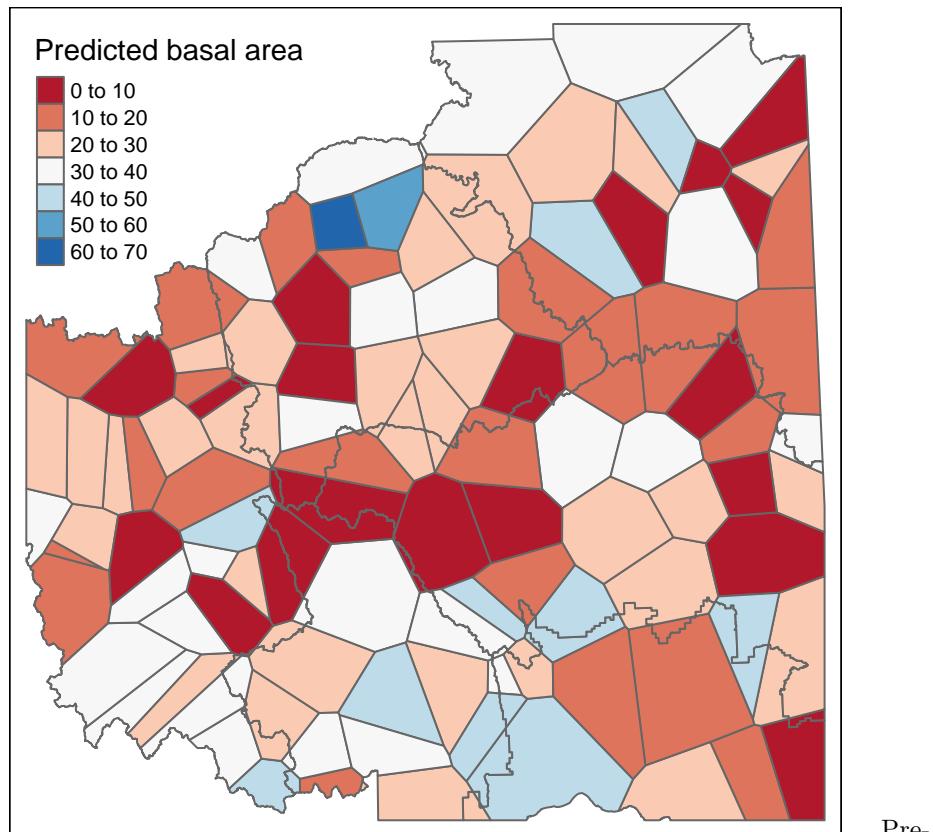
### Step 3: The tessellated surface does not store attribute information from the point data layer. Hence, the information from the data layer should be carried over to tessellated surface

```
#We'll use the over() function from the "sp" package to join the point attributes to the thiessian polygons
th.z      <- over(th,dta, fn=mean)
th.spdf   <- SpatialPolygonsDataFrame(th, th.z)

# Finally, we'll clip the tessellated surface to the Texas boundaries
th.clp    <- raster::intersect(cata,th.spdf)
```

#### Step 4: Visualize the results

```
fig_cap <- paste0("Predicted basal area over the entire Fort Saint Johns TSA using thiessen polygon")  
# Map the data  
#using package "tmap"  
tm_shape(th.clp) +  
  tm_polygons(col="baha_L", palette="RdBu", auto.palette.mapping = FALSE, title="Predicted basal area")  
  tm_legend(legend.outside=FALSE)
```



Predicted basal area over the entire Fort Saint Johns TSA using thiessen polygon.

#### Inverse Distance Weighting

\***Inverse distance weighting (IWD)** [Shepard, 1968] estimates the variable of interest by assigning more weight to closer points using the weighting function ( $w$ ) based on the weighting exponent known as power ( $p$ ) [Babak and Deutsch, 2009]. The influence of one data point on the other decreases as the distance increases. Hence, higher power of the exponent will result in point of interest having less effect on the points far from it [Babak and Deutsch, 2009]. It is a simple technique that does not require prior information to be applied to spatial

prediction [Shepard, 1968]. Lower value of exponents mean more averaging, and the weights are more evenly distributed among the surrounding data points [Shepard, 1968].

#### Step 1: Fix the projections between data points and shape file

```
# project the data based on the colorado plateau boundry projection

dsp <- SpatialPoints(plot1[,14:15], proj4string=CRS("+proj=utm +zone=10 +ellps=GRS80 +datum=NAD83"))
# convert the data into spatial object
dsp <- SpatialPointsDataFrame(dsp, plot1)

TA <- CRS("+proj=utm +zone=10 +ellps=GRS80 +datum=NAD83")
library(rgdal)
dta <- spTransform(dsp, TA)
cata <- spTransform(spdf, TA)
```

#### Step 2: Create empty grid or over-lay empty grid over the study area

An empty grid over Fort Saint Johns TSA is created, where n is the total number of cells over which interpolation is to be done.

```
grd           <- as.data.frame(spsample(dta, "regular", n=5000))
names(grd)     <- c("X", "Y")
coordinates(grd) <- c("X", "Y")
gridded(grd)   <- TRUE # Create SpatialPixel object
fullgrid(grd)  <- TRUE # Create SpatialGrid object

# Add P's projection information to the empty grid
proj4string(dta) <- proj4string(dta) # Temp fix until new proj env is adopted
proj4string(grd) <- proj4string(dta)
```

#### Step 3: Interpolate the grid cells using a power value of 2

Power values can be adjusted depending on characteristics of variable being interpolated

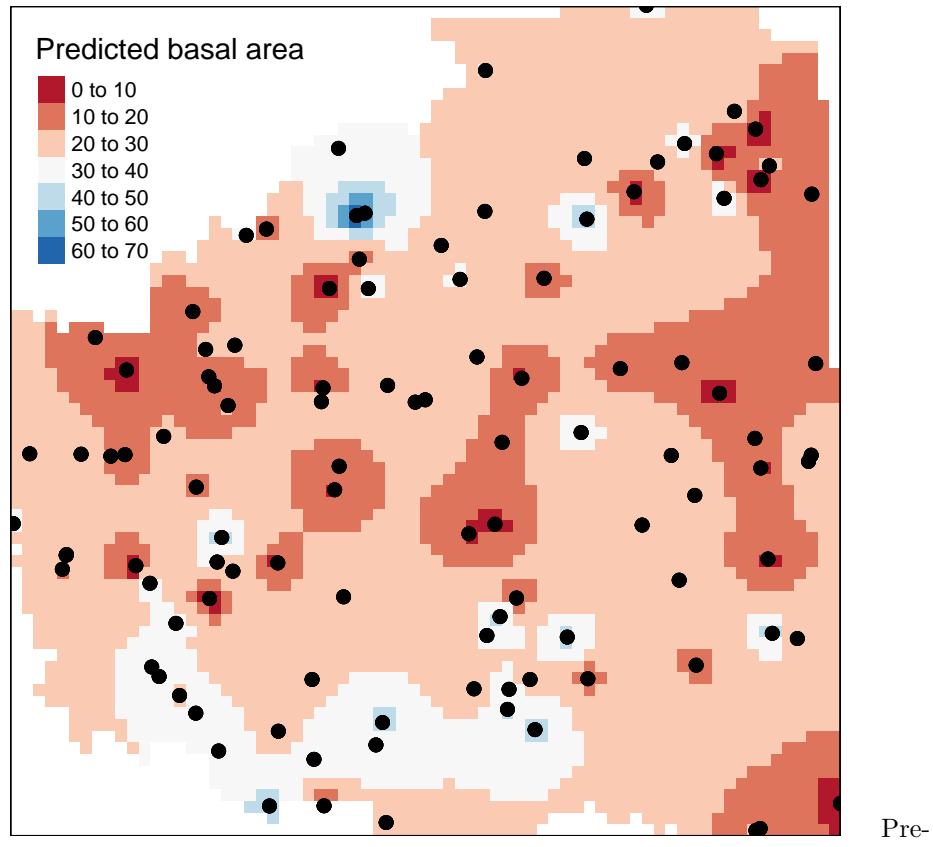
```
P.idw <- gstat::idw(baha_L ~ 1, dta, newdata=grd, idp=2.0)

## [inverse distance weighted interpolation]
# Convert to raster object then clip to Texas
r       <- raster(P.idw)
r.m    <- mask(r, cata)
```

#### Step 4: Plot the results from inverse distance weighting interpolation using

```
fig_cap <- paste0("Predicted basal area over the entire Fort Saint Johns TSA using inverse distance weighting interpolation using a power value of 2.0")
tm_shape(r.m) +
  tm_raster(n=8, palette = "RdBu", auto.palette.mapping = FALSE,
```

```
title="Predicted basal area") +
tm_shape(dta) + tm_dots(size=0.2) +
tm_legend(legend.outside=FALSE)
```



Predicted basal area over the entire Fort Saint Johns TSA using inverse distance weighting.

#### Step 5: Leaflet map for some interaction

Predicted basal area over the entire Fort Saint Johns TSA using inverse distance weighting projected over the province of British Columbia. Nepal, CC-BY-SA 4.0.

### 10.37 Methods Using Semi-variogram

### 10.38 Kriging

The spatial interpolation technique such as inverse distance weighting(IWD), nearest neighbor, and polygon approach are based on the surrounding neigh-

borhood. There is another group of interpolation methods generally known as **kriging** [Krige, 1951] which is based on both surrounding neighborhood and statistical models, especially **spatial autocorrelation**. Kriging uses the variogram modeling approach we studied in section 10.3 as a statistical model and incorporates the information about **spatial autocorrelation** while performing the interpolation. Since kriging uses the geostatistical model it has capacity of both prediction and provides some measure of the accuracy of prediction [Goovaerts, 2008]. The basic assumption of kriging is that the distance based samples reflect some degree of spatial correlation [Goovaerts, 2008]. We should note one thing that kriging works with raster surfaces where variable of interest are to be interpolated using the sampled locations. Kriging works with the following basic mathematical model:

$$\hat{Z}_{s_0} = \sum_i^N \lambda_i Z_{s_i}$$

Where,

$\hat{Z}_{s_0}$  = variable of interest to predicted at unsampled location  $s_0$

$\lambda$  = an unknown value of weight at the measured  $s_i$  location

$Z_{s_i}$  = measured value at the sampled location  $s_i$

The goal of kriging is to determine the weights

$$\lambda_i$$

that will minimize the variance estimator of the predicted value and actual value at the unsampled location:

$$Var|\hat{Z}_{s_0} - Z_{s_0}|$$

The

$$\hat{Z}_{s_0}$$

is decomposed into a trend component

$$\mu_{s_0}$$

, which is the mean function as seen in the following equation:

$$\hat{Z}_{s_0} = \mu_{s_0} + \epsilon_{s_0}$$

Where,

$\epsilon_{s_0}$  = spatially autocorrelated errors

### 10.38.1 Linear Kriging

Linear kriging are distribution free linear interpolation techniques that are in alignment with linear regression methods [Asa et al., 2012]. There are three principle linear kriging techniques as discussed below:

## 10.39 Case Study: Title of Case Study here

You see textual case study content here

For this case study, we will use ground plot data from Young stand monitoring (YSM) program data [Province of BC, 2018] for Fort Saint Johns timber supply area (TSA) in the province of British Columbia, Canada. Fort Saint Johns is divided into 6 blocks respectively **Figure 10.11**. There is a total of 108 YSM plots used in this study **Figure 10.11**. The total basal area ( $\text{m}^2/\text{ha}$ ) is our variable of interest in this study. For each of the YSM plot the total basal area was calculated by adding the basal area for all trees within the plot. We will use this dataset to explore different interpolation technique to find the variable of interest (basal area) in the unsampled locations.

::::

## 10.40 Simple Kriging

Simple kriging works with the assumption that the mean is known and constant over entire domain and calculated as the average of the data [Wackernagel, 2002]. The number of sampled points used to make the prediction of the variable of interest in unmeasured location depends upon the range of semivariogram model used [Burrough and McDonnell, 1998].

$$\hat{Z}_{s_o} = \mu_{s_o} + \epsilon_{s_o}$$

Where,

$\hat{Z}_{s_o}$  = variable of interest predicted at a given spatial location  $s_o$

$\mu_{s_o}$  = an known constant mean

**Step 1: Make sure the projection of the point data and shape file is same**

```
# summarize the data to individual plots
data<-plot1 %>%
  group_by(utm_eastin,utm_northi) %>%
  summarise(total= sum(baha_L))
# convert the data to spatial point data frame and change the projection to NAD83
dsp <- SpatialPoints(data[,1:2], proj4string=CRS("proj=utm +zone=10 +ellps=GRS80 +dat
```

```

dsp <- SpatialPointsDataFrame(dsp, data)

# Make the projection similar for plot and polygon data
TA <- CRS("+proj=utm +zone=10 +ellps=GRS80 +datum=NAD83")
dta <- spTransform(dsp, TA)
cata <- spTransform(spdf, TA)

```

### Step 2: Create an empty grid

A grid with the total number of n cells where basal area is to be predicted is overlaid over Forty Saint Johns. The grid will be a raster

```

grd           <- as.data.frame(spsample(dta, "regular", n=10000))
names(grd)    <- c("X", "Y")
coordinates(grd) <- c("X", "Y")
gridded(grd)  <- TRUE # Create SpatialPixel object
fullgrid(grd) <- TRUE # Create SpatialGrid object

# Add projection information to the empty grid relative to the plot and polygon projection
proj4string(dta) <- proj4string(dta) # Temp fix until new proj env is adopted
proj4string(grd) <- proj4string(dta)

```

### Step 3: Calculate the overall mean of the variable to be interpolated

Simple kriging, which assumes that mean is a known constant over entire domain need a mean value of variable of interest (basal area)

```

basal<-mean(data$total)
basal

```

```
## [1] 95.72555
```

### Step 4: Semivariogram modeling

Start by fitting the semivariogram model for the variable of interest (basal area) and see which model best fit the data

```

fig_cap <- paste0("Variogram models fitted for basal area using the YSM plot data. Nepal, CC-BY-S
##### Exponential variogram #####
TheVariogram=variogram(total~1, data=dta)
TheVariogramModel <- vgm(psill=2500, model="Exp", nugget=1500, range=20000)
FittedModel <- fit.variogram(TheVariogram, model=TheVariogramModel)
preds = variogramLine(FittedModel, maxdist = max(TheVariogram$dist))
g<-ggplot(TheVariogram,aes(x=dist,y=gamma))+geom_point()+
  geom_line(data = preds)+ theme_classic()+
  labs(x = "lag distance (h)", y = "Semi-variance")+
  theme_bw()+ ggtitle("Exponential")+
  theme(text = element_text(size =14))+
  theme(axis.title.x=element_blank(),
```

```

    axis.title.y=element_text(size=14),
    axis.text.x =element_blank(),
    axis.text.y =element_text(size=14))

##### Spherical Variogram
TheVariogramModel1 <- vgm(psill=2500, model="Sph", nugget=1500, range=20000)
FittedModel1 <- fit.variogram(TheVariogram, model=TheVariogramModel1)
preds1 = variogramLine(FittedModel1, maxdist = max(TheVariogram$dist))
h<-ggplot(TheVariogram,aes(x=dist,y=gamma))+geom_point()+
  geom_line(data = preds1)+ theme_classic()+
  labs(x = "lag distance (h)", y = "Semi-variance")+
  theme_bw()+ggtitle("Spherical")+
  theme(text = element_text(size =14))+
  theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.x =element_blank(),
        axis.text.y =element_blank(),
        axis.ticks.y= element_blank())

## Gaussian Variogram
TheVariogramModel2 <- vgm(psill=2500, model="Gau", nugget=1500, range=20000)
FittedModel2 <- fit.variogram(TheVariogram, model=TheVariogramModel2)
preds2 = variogramLine(FittedModel2, maxdist = max(TheVariogram$dist))
i<-ggplot(TheVariogram,aes(x=dist,y=gamma))+geom_point()+
  geom_line(data = preds2)+ theme_classic()+
  labs(x = "lag distance (h)", y = "Semi-variance")+
  theme_bw()+ggtitle("Gaussian")+
  theme(text = element_text(size =14))+
  theme(axis.title.x=element_text(size=14),
        axis.title.y=element_text(size=14),
        axis.text.x =element_text(size =14),
        axis.text.y =element_text(size=14))

## circular Variogram
TheVariogramModel3 <- vgm(psill=2500, model="Cir", nugget=1500, range=20000)
FittedModel3 <- fit.variogram(TheVariogram, model=TheVariogramModel3)
preds3 = variogramLine(FittedModel3, maxdist = max(TheVariogram$dist))
j<-ggplot(TheVariogram,aes(x=dist,y=gamma))+geom_point()+
  geom_line(data = preds3)+ theme_classic()+
  labs(x = "lag distance (h)", y = "Semi-variance")+
  theme_bw()+ggtitle("Circular")+
  theme(text = element_text(size =14))+
  theme(axis.title.x=element_text(size=14),
        axis.title.y=element_blank(),
        axis.ticks = element_blank(),
        axis.text.y =element_blank())

```

```

axis.text.x =element_text(size =14),
axis.text.y =element_blank()

#####
##### combine all the plots together #####
#####
grids_bs <- plot_grid(g,h,i,j,ncol=2,align = "h")
grids_bs

```

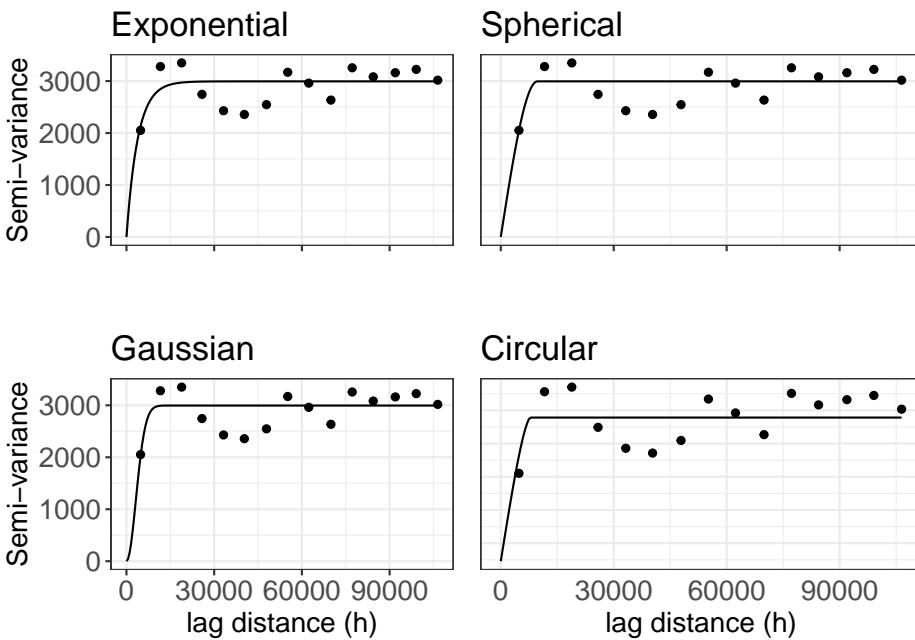


Figure 10.20: Variogram models fitted for basal area using the YSM plot data. Nepal, CC-BY-SA-4.0.

Variogram models fitted for basal area using the YSM plot data.

**Step 5:** Put all the variogram parameters in a table and see which fits best

**Table 10.2** Summary of various components of variogram fro four variogram models.

Model	Range	Nugget	Sill	Partial_sill	Partial_sill_to_Sill
Exponential	3872.10	0.00	2994.10	2994.10	1
Spherical	9704.83	0.00	2994.64	2994.64	1
Gaussian	4552.88	0.00	2995.03	2995.03	1
Circular	7801.57	729.25	2161.74	1432.49	0.66

Looking into the variogram and the parameter (Figure 10.20, Table: , we can

see that exponential variogram fits the data quite well compared to spherical and Gaussian as it has short range, low nugget and high partial sill to total sill ratio. As pointed earlier, the spatial autocorrelation only disappear at infinite lag using exponential model and best way to go is Gaussian for our data.

**Step 6:** We will use  $\beta = 95.73$ , as we know the assume mean is a known constant over the domain for simple kriging

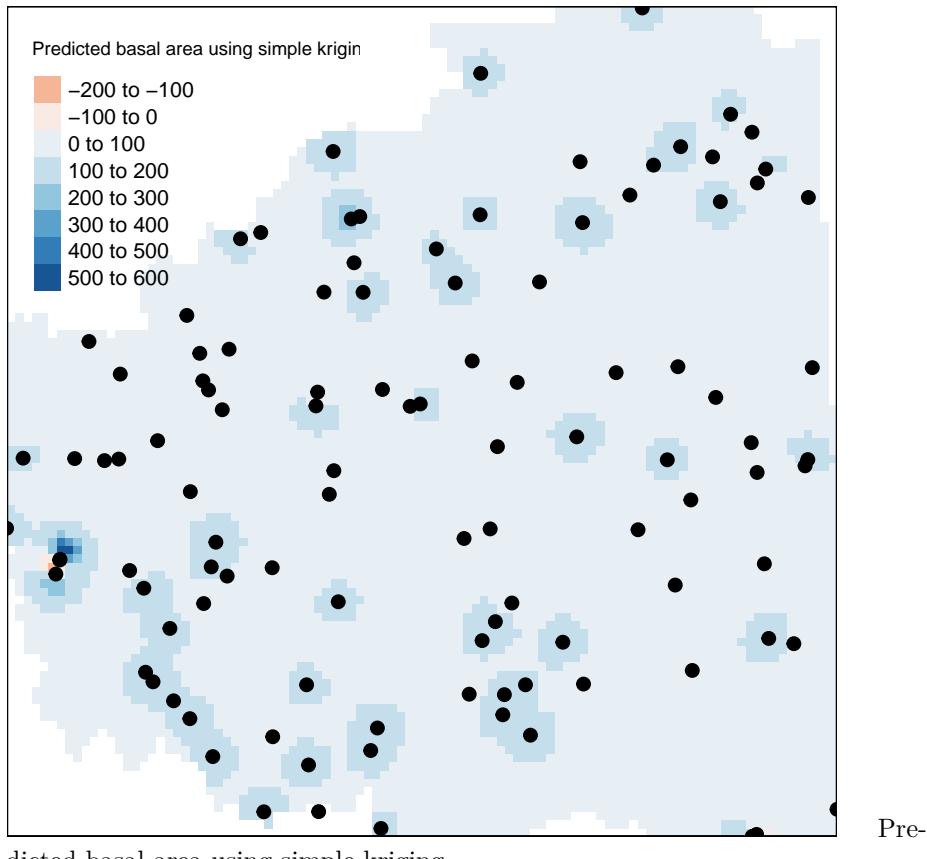
```
simple<- krige(total ~ 1, dta, grd, model=FittedModel2, beta=95.73)
```

```
## [using simple kriging]
```

**Step 6:** Visualize the predicted surface

```
fig_cap <- paste0("Predicted basal area using simple kriging. Nepal, CC-BY-SA-4.0.")
#convert the kriging results to raster
raster_krig      <- raster(simple)
raster_clip      <- mask(raster_krig, cata)

# Plot the kriging result using the library tmap
tm_shape(raster_clip) +
  tm_raster(n=8, palette = "RdBu", auto.palette.mapping = FALSE,
            title="Predicted basal area using simple kriging") +
  tm_shape(dta) + tm_dots(size=0.2) +
  tm_legend(legend.outside=FALSE)
```



dicted basal area using simple kriging.

#### Step 7: Cross validation

```
cv_sK <- krige.cv(total ~ 1, dta, model=FittedModel2, nfold=nrow(dta),
verbose=FALSE)
### calculate RMSE
res <- as.data.frame(cv_sK)$residual
sqrt(mean(res^2))

## [1] 60.42301
##### Mean residual
mean(res)

## [1] -0.6926475
##### Mean squared deviation of the prediction VS the sample
mean(res^2/as.data.frame(cv_sK)$var1.var)

## [1] 5.258013
```

## 10.41 Ordinary Kriging

Ordinary kriging [Matheron, 1973] is one of the most widely used method of kriging which assume that the mean for variable of interest is an unknown constant within the domain. The mean is calculated based on the sample that is within the search window, i.e., local mean instead of assumed constant mean over entire domain [Clark and Harper, 2007] [Goovaerts, 2008]. It assumes the following model:

$$\hat{Z}_{s_o} = \mu_{s_o} + \epsilon_{s_o}$$

Where,

$\hat{Z}_{s_o}$  = variable of interest predicted at a given spatial location  $s_o$

$\mu_{s_o}$  = an unknown constant mean

Steps 1, 3, 4, are similar to what we did for simple kriging and we don't need step 2 as mean is assumed to be a unknown constant

### Step 5: Using ordinary kriging

Kriging model is specified only using variable of interest.

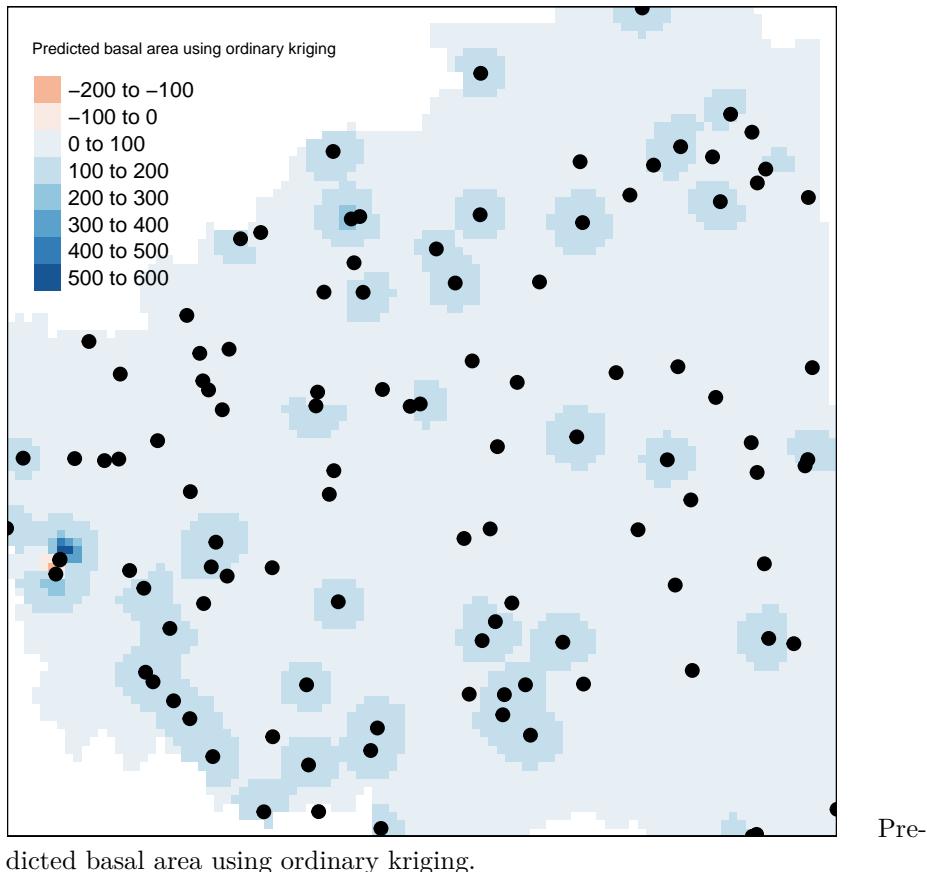
```
ordinary <- krige (total ~ 1, dta, grd, model=FittedModel2)
```

```
## [using ordinary kriging]
```

### Step 6: Visualize the predicted surface

```
fig_cap <- paste0("Predicted basal area using ordinary kriging. Nepal, CC-BY-SA-4.0.")
#convert the kriging results to raster
raster_0k      <- raster(ordinary)
Ok_clip        <- mask(raster_0k, cata)

# Plot the kriging result using the library tmap
tm_shape(Ok_clip) +
  tm_raster(n=8,palette = "RdBu", auto.palette.mapping = FALSE,
            title="Predicted basal area using ordinary kriging") +
  tm_shape(dta) + tm_dots(size=0.2) +
  tm_legend(legend.outside=FALSE)
```



#### Step 7: Cross validation

```
cv_oK <- krige.cv(total ~ 1, dta, model=FittedModel2, nfold=nrow(dta),
verbose=FALSE)
### calculate RMSE
res <- as.data.frame(cv_oK)$residual
sqrt(mean(res^2))

## [1] 60.42301
##### Mean residual
mean(res)

## [1] -0.6926475
##### Mean squared deviation of the prediction VS the sample
mean(res^2/as.data.frame(cv_oK)$var1.var)

## [1] 5.258013
```

## 10.42 Universal Kriging

The universal kriging is one of the variant of ordinary kriging. This method assume that the mean varies from location to location in a deterministic way (trend or drift) while the variance is constant throughout the domain [Matheson, 1962]. One example could be measurements of temperatures, which are commonly related to elevation (at a known rate of oC by m difference).

**Step 1:** This is similar to what we did for simple kriging and we will calculate the mean based on the drift or trend in data (more localized mean) using the spatial location of the plots

```
# Add X and Y to our original point dataframe, it is just how universal kriging formula
dta$X <- coordinates(dta) [,1]
dta$Y<- coordinates(dta) [,2]

## We will model the trend or drift using the location as X and y
TheVariogram4=variogram(total~X+Y, data=dta)
TheVariogramModel4 <- vgm(psill=2500, model="Gau", nugget=1500, range=20000)
FittedModel4 <- fit.variogram(TheVariogram4, model=TheVariogramModel4)
preds4 = variogramLine(FittedModel4, maxdist = max(TheVariogram4$dist))

g<-ggplot(TheVariogram4,aes(x=dist,y=gamma))+geom_point()+
  geom_line(data = preds4)+ theme_classic()+
  labs(x = "lag distance (h)", y = "Semi-variance")+
  theme_bw()+ ggtitle("Gaussian")+
  theme(text = element_text(size =14))+
  theme(axis.title.x=element_blank(),
        axis.title.y=element_text(size=14),
        axis.text.x =element_blank(),
        axis.text.y =element_text(size=14))

## using locations of the sample to calculate the localized mean
universal<- krige(total~X+Y,dta, grd, model=FittedModel4)
```

## [using universal kriging]

**Step 2:** Visualize the results as we usually

```
fig_cap <- paste0("Predicted basal area using universal kriging. Nepal, CC-BY-SA-4.0.")
#convert the kriging results to raster
raster_uk      <- raster(universal)
uk_clip       <- mask(raster_uk, cata)

# Plot the kriging result using the library tmap
tm_shape(uk_clip) +
  tm_raster(n=8,palette = "RdBu", auto.palette.mapping = FALSE,
            title="Predicted basal area using universal kriging") +
```

```
tm_shape(dta) + tm_dots(size=0.2) +
tm_legend(legend.outside=FALSE)
```

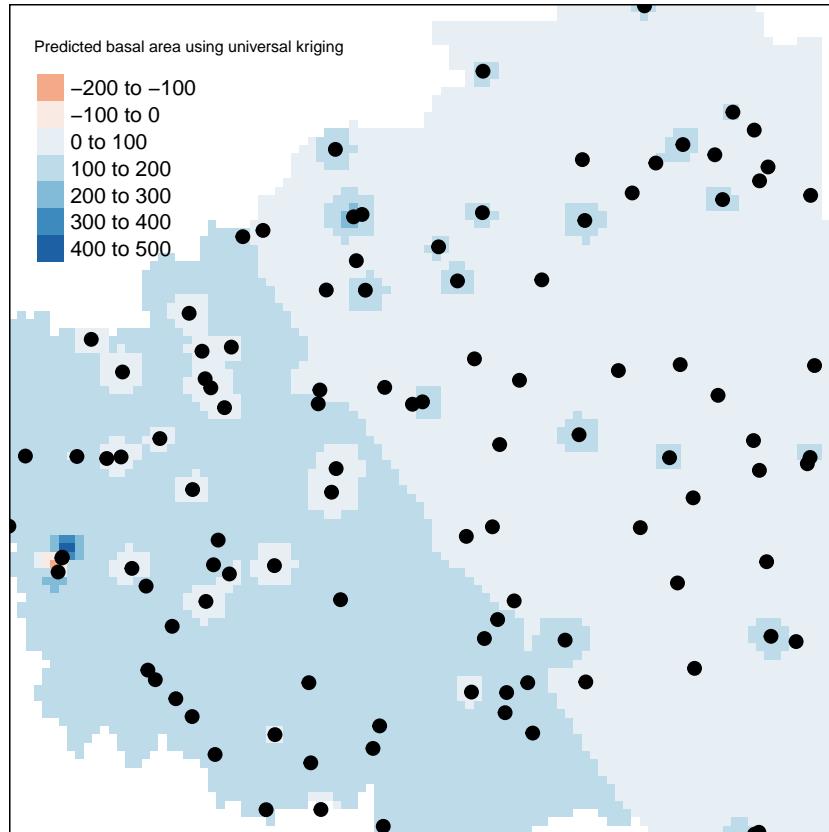


Figure 10.21: Predicted basal area using universal kriging. Nepal, CC-BY-SA-4.0.

Predicted basal area using universal kriging.

### Step 3: Cross validation

```
cv_uK <- krige.cv(total~ X+Y, dta, model=FittedModel4, nfold=nrow(dta),
verbose=FALSE)
### calculate RMSE
res <- as.data.frame(cv_uK)$residual
sqrt(mean(res^2))

## [1] 57.80639
#### Mean residual
mean(res)
```

```

## [1] -0.3187457
##### Mean squared deviation of the prediction VS the sample
mean(res^2/as.data.frame(cv_uK)$var1.var)

## [1] 4.048899

```

## Which method is the best given our data?

We will put the cross validation results together and cross compare across three methods.

**Table 10.3** Cross-validation results for different type of kriging.

Method	RMSE	ME	MSDR
Simple	60.42	-0.69	5.25
Ordinary	60.42	-0.69	5.25
Universal	68.48	-1.37	10.97

From the cross validation results (**Table 10.3**), we want root mean squared error (RMSE) to be low for greater predictive accuracy [Tziachris et al., 2017]. We also want mean error (ME) to be as close to 0 as possible [Tziachris et al., 2017]. And we want mean squared deviation Ratio (MSDR) to be closer to 1 for the good kriging model [Tziachris et al., 2017]. The RMSE is lowest for simple and ordinary kriging, the ME is negative and below zero, and the MSDR of the predictions vs. the sample is low and closer to 1 for both simple and ordinary kriging (Table 10.3). This means the variability in predictions from both kriging are somewhat closer to real values than universal kriging [Tziachris et al., 2017]. Looks like universal kriging is predicting more negative basal area for some of the location within TSA. The choice between simple and ordinary kriging may vary with researchers discretion here.

## 10.43 Co-Kriging

Co-kriging uses the information about different covariates to predict the value of the variable of interest at an unsampled location [Cressie, 1994]. It utilizes the autocorrelation of the variable of interest and the cross correlations between the variable of interest with all the covariates to make the prediction [Cressie, 1994]. In order to implement co-kriging we need to have a strong correlation between the covariates [Tziachris et al., 2017]. The spatial variability of one variable should be correlated with the spatial variability of the other covariates [Tziachris et al., 2017].

## 10.44 Non-Linear Kriging

In principle, nonlinear kriging algorithms are linear kriging algorithms applied to nonlinear transformations of the data points into a continuous variable [Deutsch and Journel, 1993]. We will briefly talk about four principal non-linear kriging techniques in this chapter while our focus mostly is on linear kriging methods.

## 10.45 Indicator Kriging

Indicator kriging is a non-parametric approach of estimating a binary variable (presence absence or variables that takes 0 or 1 value) of interest at an unsampled or unmeasured location [Journel, 1983]. For example, we might have a sample that consists of information on presence or absence of Douglas-fir tree species within Williams lake timber supply area, where 0 indicates absence and 1 indicates the presence of species. Indicator kriging assumes that mean is a unknown constant over the domain. The only difference between the indicator kriging and ordinary kriging is in the use of binary variable. The basic mathematical formulation of indicator kriging is given below:

$$I_s = \mu + \epsilon_s$$

where,

$I$  = binary variable preicated at the location  $s$

$\mu$  = unknown mean

$\epsilon_s$  = spatially autocorrelated error

## 10.46 Probability Kriging

Probability kriging is useful when the variable of interest is binary as in case of indicator kriging. It is a special form of co-kriging which estimate the conditional probability that the unknown value of a variable at an unsampled location is above a specified cutoff level [Carr and hsien Mao, 1993]. As in co-kriging, this method utilizes the autocorrelation of variable of interest and the cross correlations between the variable of interest with all the covariates to make the prediction [Carr and hsien Mao, 1993]

## 10.47 Disjunctive Kriging

Disjunctive kriging allows to estimate the value of a variable of interest at an unsampled location and estimating the conditional probability that the unknown value of a variable at an unsampled location is above a specified cutoff level [Yates et al., 1986]. Disjunctive kriging transforms the data into a normal distribution and then determine the probability that true value of variable at each location of interest exceeds the predefined threshold or cut-off probability [Daya and Bejari, 2015].

## 10.48 Spatial Regression Models

For classical statistics tests, **spatial autocorrelation** is problematic as ordinary least square (regression) or analysis of variance (ANOVA) assumes that observations are independent in space and time [Meng et al., 2009]. However, geostatistical data violates the assumptions of independence, and using regression and ANOVA might inflate the significance of t and F statistics, when, in fact, they may not be significant at all [Meng et al., 2009]. In that case one should try to improve the regression model by adding important **auxiliary** (independent variables that are associated or important in predicting the variables of interest) and incorporating the spatial autocorrelation structure [Meng et al., 2009] using spatial regression models [Anselin and Bera, 1998]. **The whole objective of spatial regression is to understand the association of the variable of interest with the independent variables while accounting for the spatial structure present in the data.** We will show two examples of spatial regression model in this section using our familiar YSM data for Fort Saint Johns TSA.

## 10.49 Case Study: Title of case study here

You see textual case study content here

For this case study, we will use ground plot data from Young stand monitoring (YSM) program data [Province of BC, 2018] for Fort Saint Johns timber supply area (TSA) in the province of British Columbia, Canada. Fort Saint Johns is divided into 6 blocks respectively **Figure 10.11**. There is a total of 108 YSM plots used in this study **Figure 10.11**. The total basal area ( $m^2/ha$ ) is our variable of interest or response variable in this study. For each of the YSM plot the total basal area was calculated by adding the basal area for all trees within the plot. We will use the auxiliary variables such as trees per hectare (TPH), elevation (m), site index, top height(m), and tree volume ( $m^3/ha$ )

:::::

## 10.50 Spatial Lag Model

Spatial lag models assume that the spatial autocorrelation only exist in the response variable or the variable of interest [Anselin and Bera, 1998]. Spatial lag model has the following general mathematical formulations:

$$y = (\rho)WY + \beta X + \epsilon$$

where,

$y$  = response variable or variable of interest

$\rho$  = coefficients for the spatial weight matrix W

$\beta$  = coefficients for the predictor variables

$WY$  = spatially lagged response variable for the weight matrix  $W$

$X$  = matrix of observations for the predictor variables

$\epsilon$  = vector of error terms

## 10.51 Steps in Fitting Spatial Lag Model:

### Step 1: Build spatial weight matrix

We will create a **spatial weight matrix** using the **distance-based approach** usign the **nearest-neighbor** approach. This spatial weight matrix will be used in our spatial lag model to account for the spatial autocorreltion.

### Step 2: Check for the spatial autocorrelation

We will check the spatial autocorrelation using the Moran's I adn Moran's test using our distance based spatial weight matrix [Getis and Aldstadt, 2010] we have just calculated. For the details on Moran's I, please refer to **section 4.3**.

```
##  
## Moran I test under randomisation  
##  
## data: YSM_pots$Basal  
## weights: YSM.W  
##  
## Moran I statistic standard deviate = 2.6524, p-value = 0.003996  
## alternative hypothesis: greater  
## sample estimates:  
## Moran I statistic      Expectation      Variance  
##       0.0656590005     -0.0093457944    0.0007996302
```

Moran's I value was ( $I=0.065$ ), which was significant ( $p\text{-value} < 0.05$ ) indiecting a positive spatial autocorrelation. It seems like a weak autocorrelation but for the purpose of demonstration in this section we will proceed further assuming there is a spatial autocorrelation.

### Step 3: Fit a spatial lag model

In this step, we will fit a spatial lag model using basal area ( $\text{m}^2/\text{ha}$ ) as a **response variable**. While, trees per hectare (TPH), elevation (m), site index, top height(m), and tree volume ( $\text{m}^3/\text{ha}$ ) will be used as **predictor variable**.

```
##  
## Call:lagsarlm(formula = f1, data = YSM_pots, listw = YSM.W, zero.policy = T)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -28.4814  -8.7285  -1.5817  8.6639  34.8502
```

```

## 
## Type: lag
## Coefficients: (numerical Hessian approximate standard errors)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.0836e+01 1.2063e+01 -1.7273 0.0841175
## TPH          3.7555e-03 4.6869e-04  8.0128 1.11e-15
## Elevation    2.0548e-02 5.4083e-03  3.7993 0.0001451
## Volume       9.6251e-02 4.9212e-03 19.5584 < 2.2e-16
## Height       -8.6348e-02 4.7398e-01 -0.1822 0.8554439
## Site_index   -4.3118e-01 3.2905e-01 -1.3104 0.1900755
##
## Rho: 0.18185, LR test value: 5.015, p-value: 0.025129
## Approximate (numerical Hessian) standard error: 0.080031
##      z-value: 2.2722, p-value: 0.023072
## Wald statistic: 5.163, p-value: 0.023072
##
## Log likelihood: -425.7183 for lag model
## ML residual variance (sigma squared): 155.08, (sigma: 12.453)
## Number of observations: 108
## Number of parameters estimated: 8
## AIC: 867.44, (AIC for lm: 870.45)

```

#### Step 4: Select auxiliary variables and refit the model

Varieties of ways has been proposed to select the auxiliary variables to get the best spatial model. We will go through them briefly. First, we will Use the alpha=0.005 to check whether our auxiliary variables are significantly associated with the basal area. From the summary, we can see that TPH, Elevation and Volume has p-value < 0.05 indicating that they are significantly associated with basal area.

```

## 
## Call:lagsarlm(formula = f2, data = YSM_pots, listw = YSM.W, zero.policy = T)
## 
## Residuals:
##      Min        1Q     Median        3Q        Max
## -29.94846 -7.71629 -0.30949  8.76752  35.90219
## 
## Type: lag
## Coefficients: (numerical Hessian approximate standard errors)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.8815e+01 8.8943e+00 -3.2397 0.001196
## TPH          3.8730e-03 4.1431e-04  9.3481 < 2.2e-16
## Elevation    2.4226e-02 4.5489e-03  5.3257 1.006e-07
## Volume       9.4008e-02 2.3600e-03 39.8342 < 2.2e-16
##
## Rho: 0.16445, LR test value: 4.1412, p-value: 0.041852

```

```

## Approximate (numerical Hessian) standard error: 0.07985
##      z-value: 2.0595, p-value: 0.039449
## Wald statistic: 4.2414, p-value: 0.039449
##
## Log likelihood: -426.8261 for lag model
## ML residual variance (sigma squared): 158.35, (sigma: 12.584)
## Number of observations: 108
## Number of parameters estimated: 6
## AIC: 865.65, (AIC for lm: 867.79)

```

**Step 5: Assess both models using Akaike Information Criteria (AIC)**

Sometime, only using the p-value to assess significant variables will not be useful while assessing which models best fits the data as multiple models can have potential to describe the association between response and predictor variables. In the context, when we have different competing models, we can use AIC [for details see, Akaike, 1973] to compare the models. For example, suppose models we have fitted in **step 3** and **step 4** were competing and potential. We can select the best model between two with lowest AIC values. AIC value of model from **step 3** is 867.44 while the AIC value of model from **step 4** is 865.65, indicating later one is the best fit to our data.

**Step 6: Interpreting rho coefficient for our selected model**

Rho (0.16445), reflects the spatial dependence inherent in our sample data, measuring the average influence on observations by their neighboring observations. It has a positive effect and it is significant (p-value < 0.05). As a result, the general model fit improved over the linear model.

## 10.52 Spatial Error Model

Spatial error models assume that the spatial autocorrelation exists in the residuals or the error term of the regression equation [Anselin and Bera, 1998]. The general mathematical formula for spatial error model is given below:

$$y = \beta X + \epsilon$$

$$\epsilon = \lambda(W)\epsilon + u$$

where,

$y$  = response variable or variable of interest

$\beta$  = coefficients for the predictor variables

$\lambda$  = coefficients for the spatial weight matrix  $W$  for spatially autocorrelated errors

$$(W)\epsilon = \text{spatial weight matrix } W$$

$X$  = matrix of observations for the predictor variables

$u$  = independent errors

### 10.53 Steps in Fitting Spatial Error Model:

**Step 1 and Step 2** in fitting spatial error model is exactly similar to what we did for spatial lag model.

#### Step 3: Fit the spatial error model

We will use the function **errorsarlm** from the package **spatialreg** to fit the error model indicating that we are accounting the spatial autocorrelation that exists in the residuals or the error terms instead of the response variable (**basal area**).

```
##  
## Call:errorsarlm(formula = f3, data = YSM_pots, listw = YSM.W, zero.policy = T)  
##  
## Residuals:  
##      Min       1Q     Median       3Q      Max  
## -27.7331 -7.8589 -1.0210  7.8231 35.1950  
##  
## Type: error  
## Coefficients: (asymptotic standard errors)  
##                  Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -0.14171644 10.59980548 -0.0134 0.98933  
## TPH          0.00378146 0.00048423  7.8092 5.773e-15  
## Elevation    0.01635979 0.00645862  2.5330 0.01131  
## Volume       0.09585359 0.00479585 19.9868 < 2.2e-16  
## Height       -0.11345603 0.45520594 -0.2492 0.80317  
## Site_index   -0.30051747 0.32907580 -0.9132 0.36113  
##  
## Lambda: 0.50793, LR test value: 3.6202, p-value: 0.057081  
## Asymptotic standard error: 0.1857  
## z-value: 2.7352, p-value: 0.0062338  
## Wald statistic: 7.4815, p-value: 0.0062338  
##  
## Log likelihood: -426.4156 for error model  
## ML residual variance (sigma squared): 154.69, (sigma: 12.438)  
## Number of observations: 108  
## Number of parameters estimated: 8  
## AIC: 868.83, (AIC for lm: 870.45)
```

#### Step 4: Select auxiliary variables and refit the model

We will use the alpha=0.05 to check whether our auxiliary variables are significantly associated with the basal area. From the summary, we can see that TPH,

Elevation and Volume has p-value < 0.05 indicating that they are significantly associated with basal area.

```
## 
## Call:errorsarlm(formula = f4, data = YSM_pots, listw = YSM.W, zero.policy = T)
## 
## Residuals:
##      Min       1Q   Median      3Q      Max
## -27.84609 -7.91303 -0.65425  7.82678 36.06371
## 
## Type: error
## Coefficients: (asymptotic standard errors)
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.12704159  6.71512837 -1.0613  0.288535
## TPH          0.00391705  0.00042822  9.1472 < 2.2e-16
## Elevation    0.01843298  0.00595547  3.0951  0.001967
## Volume       0.09366580  0.00235054 39.8487 < 2.2e-16
## 
## Lambda: 0.52491, LR test value: 3.7554, p-value: 0.052636
## Asymptotic standard error: 0.18121
## z-value: 2.8967, p-value: 0.0037711
## Wald statistic: 8.3909, p-value: 0.0037711
## 
## Log likelihood: -427.019 for error model
## ML residual variance (sigma squared): 156.21, (sigma: 12.498)
## Number of observations: 108
## Number of parameters estimated: 6
## AIC: 866.04, (AIC for lm: 867.79)
```

#### Step 5: Assess both models using akiekie information criteria (AIC)

AIC value of model from **step 3** is 868.83 while the AIC value of model from **step 4** is 866.04, indicating later one is the best fit to our data.

#### Step 6: Interpret lambda parameter from the summary from our selected model

The lag error parameter Lambda for the model in **step 4** is positive and significant (p-value < 0.05), indicating the need to control for spatial autocorrelation in the error

## 10.54 Selection Between Lag and Error Model

When it is not so clear theoretically that either of the spatial model works for our data, we can compare the model performance parameters: the AIC and Log likelihood. In our case, the spatial error model has lowest AIC and highest negative Log likelihood values. Hence, spatial lag model best fits our data.

## **Remember This?**

When the question is which of the two models is better? This is an open question. The general advice is first to look for a theoretical basis to inform your choice. If there are strong substantive grounds for one model instead of the other, you should adopt it.

## **Reflection Questions**

1. Explain probability and non-probability sampling.
2. Define spatial autocorrelation and semivariogram.
3. When do you use spatial contiguity vs. nearest neighbor?
4. What is spatial regression mostly used for? Estimation or Prediction?

## Chapter 11

# Fundamentals of Remote Sensing

Written by Paul Hacker and Paul Pickell

At some point in your life you may have wondered why the sky is blue. You may have noticed that two leaves on the same tree are slightly different shades of green. It would be perfectly natural to wonder these things and simply allow the questions to remain unanswered. After all, they likely carry minimal significance compared to the other queries of your life. What if, however, these questions could not only be answered, but also lead you to profound insights relating to your environment? What if differences in leaf color indicated an early summer drought or the initial stages of a pest outbreak that would wreak havoc on the economy? *Remote sensing* is the overarching term that refers any scientific exploration that seeks to address these, and many other questions.

### Learning Objectives

1. Understand key principles underpinning remote sensing science
2. Become familiar with specific types of energy used in remote sensing
3. Define key interactions between energy and surface materials that enable remote sensing
4. Comprehend various considerations that effect the use of remote sensing

### Key Terms

Electromagnetic, Energy, Photons, Pixel, Radiation, Resolution, Spectrum, Wavelength

## 11.1 What is Remote Sensing?

Simply put, remote sensing is any method of gathering information about an object, or objects, without physical contact. Over the course of human history, a variety of remote sensing techniques have been used. In fact, one could argue that any organism capable of observing **electromagnetic radiation (EMR)** has a built in optical remote sensing system, such as human vision. Similar arguments could be made for other senses, such as smell or hearing, but this chapter will focus strictly on techniques that capture and record electromagnetic radiation.

One of the first recorded conceptualizations of remote sensing was presented by Plato in the Allegory of a Cave, where he philosophized that the sense of sight is simply a contracted version of reality from which the observer can interpret facts presented through transient images created by light. Over the next few centuries a variety of photosensitive chemicals were discovered which enabled the transient images humans see to be recorded. This technology was called photography (see *A History of Photography*). The ability to record the interaction of light and specific objects within a scene proved enabled the preservation of information in an accessible medium. Eventually, photography became a prominent means of immortalizing everything from individual humans to exotic landscapes. After all, a picture says a thousand words.

In 1858, an enthusiastic Frenchman named Gaspard Tournachon mounted a camera on a hot air balloon and captured images of the earth below him. Eventually, Tournachon used his balloon method to capture images of Paris (Figure 11.1). For the first time it was possible to examine the distribution of buildings, fields, forests and roads across the landscape. With this, airborne remote sensing was born. Remote sensing technologies continued to advance throughout the 19th and 20th centuries, with major socio-political conflicts like World War I and II acting as crucibles for innovation. The advancement of remote sensing has continued into the 21st century and is unlikely to slow down in the future. This is due to the relevance of three key aspects.

First and foremost, remote sensing enables the observation of objects in, or from, locations that are otherwise inaccessible for humans. The observation of Mars' surface from an orbiting satellite is a one current example. A second aspect that makes remote sensing so useful is the collection of information over a large area. For example, airborne remote sensing technologies enable observations of land cover across Canada (Figure 11.2). The ability to evaluate inaccessible objects or large areas over time is a third valuable aspect of remote sensing and is particularly relevant for land management, as predictions can be informed through the observation of historic patterns and processes. This is especially true for projects aiming to restore degraded ecosystems or plan sustainable land use practices. Before exploring the designs of specific sensor or their applications, however, it is essential to grasp some key components that underpin remote sensing science.



Figure 11.1: Images of the Arc de Triomphe in Paris, France taken by Tournachone from a balloon in 1968 [Nadar, 1868].

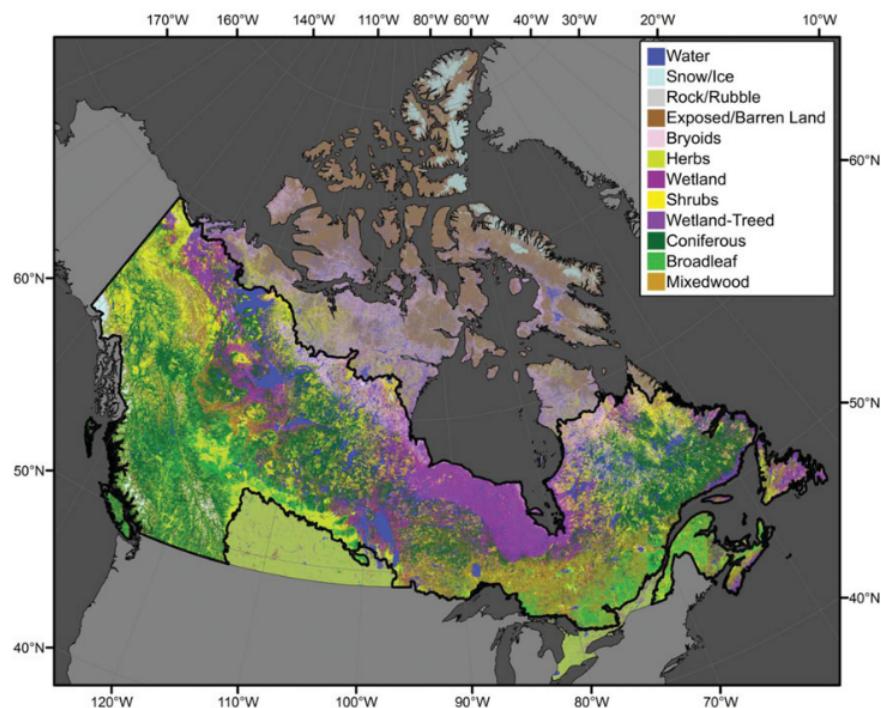


Figure 11.2: Landcover classification of Canada generated by Hermosilla et al. for the year 2005 [Hermosilla et al., 2018].

## 11.2 Measuring Energy

### 11.3 Introduction

At it's core, remote sensing is simply the measurement of photons. Although simple, this description captures the essence of remote sensing as the capacity to observe objects in our universe, from a single leaf to a distant star. You see, if an object exists it has a temperature, and any object that has a temperature emits photons. The properties of those photons are determined by how hot the object is. To really understand remote sensing, then, we must first understand photons, or electromagnetic radiation.

Essentially, photons are the smallest physical property in the electromagnetic field. Photons can be emitted from objects engaged in nuclear processes (such as the sun), objects excited thermally (like a light bulb) or objects that reflect or emit absorbed radiation. The interactions between emitted photons and other particles can be observed and used to evaluate the properties of the object. A fundamental component of a photon is it's **wavelength**, defined as the measured space between two consecutive peaks of a wave. The wavelength of a photon determines if and how it will interact with the particles around it, as well as defines the amount of **energy** it has. Measuring the differences in photon energy before and after interacting with another particle is the core of any remote sensing utilizing EMR. Equation (11.1) defines the relationship between a photon's energy and wavelength.

$$E = hc/\lambda \quad (11.1)$$

Where E is the energy of a photon, h is Planck's constant, c is the speed of light ( $c = 3 \times 10^8 \text{ m} \cdot \text{s}^{-1}$ ) and  $\lambda$  is the wavelength of the radiation. This equation contains more variables, but incorporates wavelength and in doing so utilizes an easy to measure ( $hc$  always equals  $1240 \text{ eV} \cdot \text{nm}^{-1}$ ) and familiar photon property. Due to the large range of wavelengths that photons can exhibit it is necessary to use a specific style of writing to describe them, called scientific notation.

## 11.4 Electromagnetic Spectrum

Now that you have an understanding of how the properties of photons can be measured and how to write them, we can begin to explore the **electromagnetic spectrum (EMS)**. The EMS is the continuum along which photons are located based on their properties (Figure 11.3). We have discussed both wavelength and frequency, which are inversely related and commonly used to describe EMR. Figure 11.3 also depicts a thermometer laying sideways, which demonstrates that as an object's temperature increases, the wavelength of the photons emitted decreases. This follows Equation (11.1), which demonstrates that photons with shorter wavelengths have higher energy. A practical example of this would be

that the majority of photons emitted from the sun (5,788 K) are around  $0.5 \times 10^{-6}$  nm, while the majority of photons emitted from the human body (~310 K) are around  $10^{-4}$  nm. These measurements are theoretical and are calculated using theoretical object, often called a blackbody that allows all energy to enter (no reflectance, hence “black”) and be absorbed (no transmission). The resulting EMR that is emitted would be generated thermally and be equal or greater than any other body at the same temperature. It is important to remember, however, that blackbodies do not exist and any real object will always emit a temperature less than a theoretical blackbody.

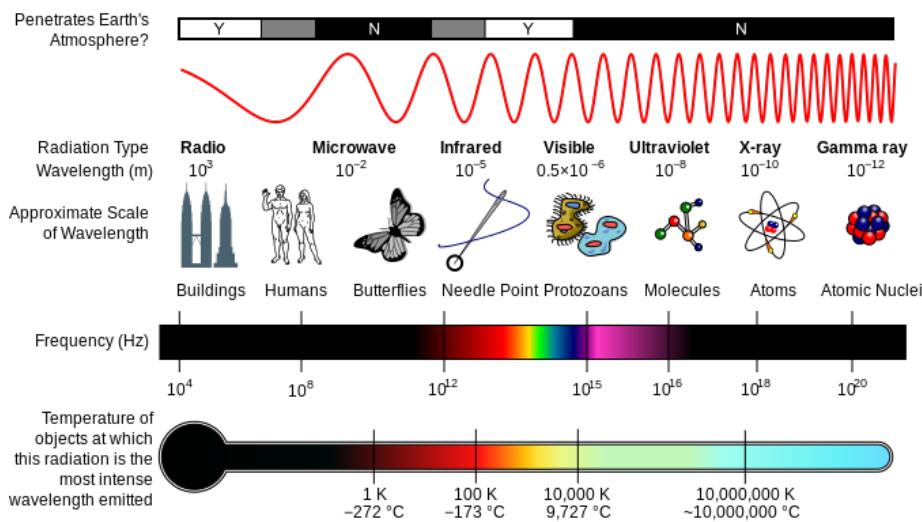


Figure 11.3: Electromagnetic (also known as Milton) spectrum depicting the type, wavelength, frequency and black body emission temperature [Inductive load and NASA, 2007].

## Call Out

Visualizing the electromagnetic spectrum (EMS) in Figure 11.3 certainly enables a wonderful comprehension of many concepts relating to photons. Perhaps more astounding, however, is the truth of how the faculty of human vision has incorporated these properties. The portion of the EMS that humans can see is between 400 nm and 750 nm, which correlates with the peak energy emitted from the sun. Perhaps it should not be surprising, but of all the possible wavelengths emitted in our environment, human eyes have evolved to maximize solar photon emission.

## 11.5 Scientific Notation

Expressing extremely large or small numbers presents a challenge to both efficiency and accessibility that has existed likely since the creation of mathematics. Scientific notation presents a simple solution to this problem through simplifying numeric presentation to a value less than 10 that is raised to a particular power. Put simply, the decimal point of a large or small number is moved to make the smallest, single digit whole number. The number of places and direction that the decimal point moves is described by an associated power of 10. Equations (11.2) and (11.3) depict how large and small numbers are presented in scientific notation, respectively.

$$1,000,000 = 1.0 \times 10^6 \quad (11.2)$$

$$0.000001 = 1.0 \times 10^{-6} \quad (11.3)$$

It is common to use mathematical operators to add, subtract, divide or multiply large numbers written in scientific notation. Addition and subtraction require the number with the smallest exponent to be altered to match the exponent of the largest number (11.4). The multiplication of two numbers written in scientific notation you simply multiply their coefficients and add their exponents (11.5). Division of two scientific numbers follows a similar structure, in which you divide the coefficients and subtract the exponents (11.6).

$$2.6 \times 10^5 + 2.0 \times 10^4 - > 2.6 \times 10^5 + 0.2 \times 10^5 = 2.8 \times 10^5 \quad (11.4)$$

$$1.5 \times 10^5 \cdot 2.0 \times 10^3 = 3.0 \times 10^8 \quad (11.5)$$

$$1.5 \times 10^5 / 2.0 \times 10^3 = 0.75 \times 10^2 = 75 \quad (11.6)$$

## 11.6 Radiation Types

Since it is possible for photon energy to vary widely across the EMS, it can be useful to group photons based on their wavelength. Generally, there are seven accepted categories. It is important to note that these categories have gradual boundaries, rather than sharp dividing lines. In order of increasing wavelength they are: radio, microwave, infrared, visible, ultraviolet, X ray and Gamma ray. We will detail each of these seven groups in Table 1 [Zwinkels, 2020]. If you wish a visual tour of the EMS you can explore “Tour of the electromagnetic spectrum” online created by Ginger Butcher for NASA in 2010.

Table 11.1: Table 1. Names and associated wavelengths for the seven regions of the electromagnetic spectrum [Zwinkels, 2020].

Name	Wavelength
Radio	1 cm - 1,000 km ( $10^3$ - $10^{10}$ )
Microwave	1 mm - 1 cm ( $10^{10}$ - $10^{11}$ )
Infrared (IR)	700 nm - 1 mm ( $10^{11}$ - $10^{14}$ )
Visible (Vis)	400 - 700 nm ( $10^{14}$ - $10^{15}$ )
Ultraviolet (UV)	10 - 400 nm ( $10^{15}$ - $10^{17}$ )
X rays	0.1 - 10 nm ( $10^{17}$ - $10^{20}$ )
Gamma rays	< 0.1 nm ( $10^{20}$ - $10^{23}$ )

## 11.7 Factors Affecting Radiation

With a solid grasp of why EMR is useful for remote sensing science and how EMR is categorized along the EMS, we can begin to apply this core knowledge with ideas and applications related to practical use. As with radiation, there are a number of key terms used to describe the fundamental concepts that make remote sensing science possible. Some of the most common terms have been included below. They are organized into three categories: Radiation Basics, Foundations of Measurement and Methods of Normalization.

## 11.8 Radiation Basics

The use of radiation to quantify properties of an object is inherently linked with relatively complex theories of physics. To minimize both confusion and workload, we will highlight a select number of key concepts that support the use of the EMS for remote sensing. The first concepts to become familiar with are radiant energy and radiant flux

Radiant energy is essentially the energy carried by photons, which is measured in Joules (J). Recall that the amount photon energy defines what wavelength (Equation (11.1)). Radiant flux, which is interchangeable with radiant power, is the amount of radiant energy that is emitted, reflected, transmitted or absorbed by an object per unit time. Radiant flux considers energy at all wavelengths and is often measured per second, making it's SI Watts (W), which is simply Joules per second ( $J \cdot s^{-1}$ ). Spectral flux is an associate of radiant flux and simply reports the amount of energy per wavelength ( $W \cdot nm^{-1}$ ) or ( $W \cdot Hz^{-1}$ ). Combined, these two terms allow us to describe the interaction with electromagnetic radiation and its environment; radiant energy interacts with an object, which results in radiant flux.

Now that you are familiar with radiant energy and flux, we can discuss irradiance. Irradiance refers to the amount of radiant energy that contacts a  $1 m^{-2}$  area

each second ( $\text{W} \cdot \text{m}^{-2}$ ). This includes all electromagnetic energy that contacts the  $1 \text{ m}^{-2}$  surface, which could be a combination of radiation from the sun, a halogen light bulb overhead and your computer screen. Another important concept is solar irradiance, which strictly refers to the amount of solar radiation interacting with our  $1 \text{ m}^{-2}$  area. Solar irradiance is very important in many remote sensing applications as it determines which photons an optical sensor *could* detect in naturally illuminated environments. An associate of irradiance is radiance, which refers to the amount of radiant flux in a specific direction. The direction in question is often called the *solid angle* and makes radiance a directional quantity. You could imagine holding a DSLR camera 90 degrees above a flat leaf so that the only item visible to the shutter is the leaf. The camera would capture the radiance reflected from the leaf's surface and the solid angle would be 90 degrees. Essentially, irradiance is used to measure the radiant energy that contacts a  $1 \text{ m}^{-2}$  area, while radiance measures the radiant flux of an object from a specific angle.

Another key idea central to understanding radiation is the theoretical blackbody. We discussed theoretical blackbodies in Section 11.4 and outlined that they are essentially an ideal body that perfectly absorbs all energy it interacts with. Blackbodies can reach thermal equilibrium and therefore can emit thermal energy in amounts directly related to their temperature. These characteristics enable the estimation of spectral emission and can be used to determine the expected spectral emission of other objects.

There are three important physical laws related to theoretical blackbodies. Plank's Law describes that an increase in a body's temperature causes both an increase in the total radiation from the body and shift in the peak of the emitted spectrum to shorter wavelengths. Wein's Law highlights the inverse relationship between temperature and wavelength and enables the estimation of the wavelength at which the peak of the emitted spectrum will occur. A third important law relevant to blackbodies is the Stefan-Boltzmann Law, which demonstrates that the temperature of a blackbody is four powers greater than the total energy it emits. Combined, these three laws demonstrate a fundamental concept of electromagnetic radiation: the temperature of a body determines how much energy is emitted and at what wavelength the spectrum peak will occur.

So far we have discussed radiant energy and flux as basic concepts interacting with a single object (leaf) or  $1 \text{ m}^{-2}$  surface. In reality, radiant energy from the sun begins interacting with objects as soon as it enters Earth's atmosphere. The process by which radiation is reflected by other particles is called scattering. Scattering occurs throughout the atmosphere and is generally separated into three categories: Rayleigh, Mie and non-selective.

The three categories of atmospheric scattering are defined by the energy's wavelength and the size of the interacting particle. When the wavelength of incoming radiation is larger than the particles (gases and water vapor) with which it interacts, Rayleigh scattering occurs. Phenomenon related to Rayleigh scattering

include Earth's sky appearing blue (Figure 11.4). When the wavelength of incoming radiation is similar to that of the particles with which it interacts Mie scattering occurs. The size of particles generally considered to be similar is between 0.1 - 10 times that of the wavelength (Figure 11.4). Smoke and dust are common causes of Mie scattering. A third type of scattering occurs when the particles involved are larger than the wavelength of the incoming radiation. This is called non-selective scattering and results in the uniform scattering of light regardless of the wavelength (Figure 11.4). Examples of non-selective scattering are clouds and fog.

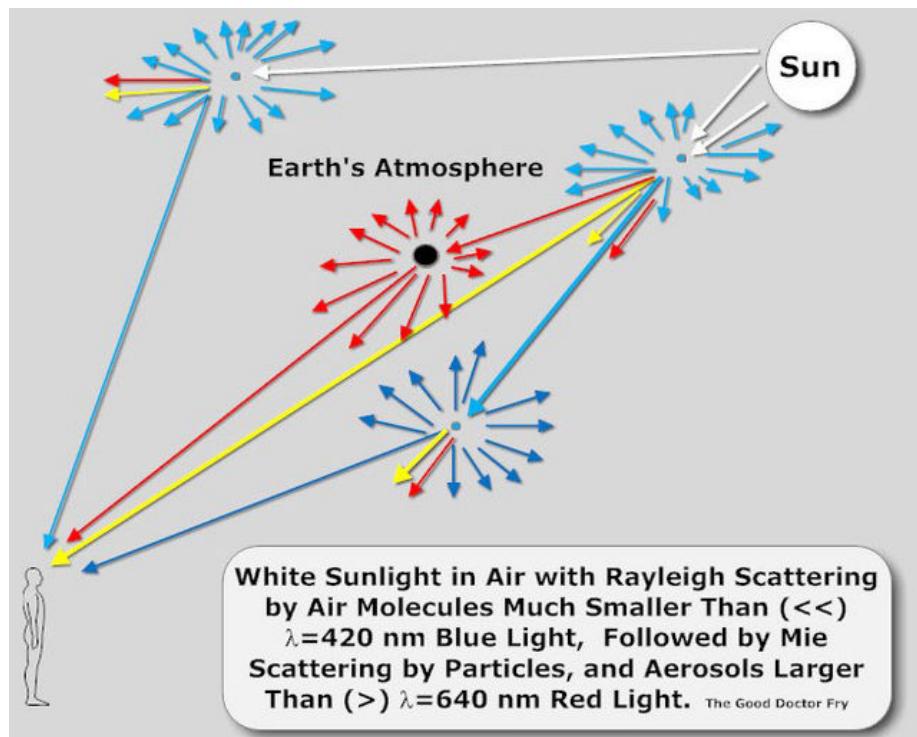


Figure 11.4: Depiction of Rayleigh, Mie and non-selective scattering [The Good Doctor Fry, 2011]. CC BY-SA 4.0.

The combination of these three scattering types leads to drastic differences between the amount of solar irradiance at the top of the atmosphere and at sea level (Figure 11.5). There are also a variety of wavelengths at which ozone, oxygen, water and carbon dioxide absorb incoming radiation, precluding entire sections of the EMS from reaching the surface. Overall, only a small portion of energy emitted from the sun reaches the Earth's surface ( $\sim 1360 \text{ W} \cdot \text{m}^{-2}$ ), depending on the time of day and year. Most energy is absorbed or scattered by particles in the Earth's atmosphere.

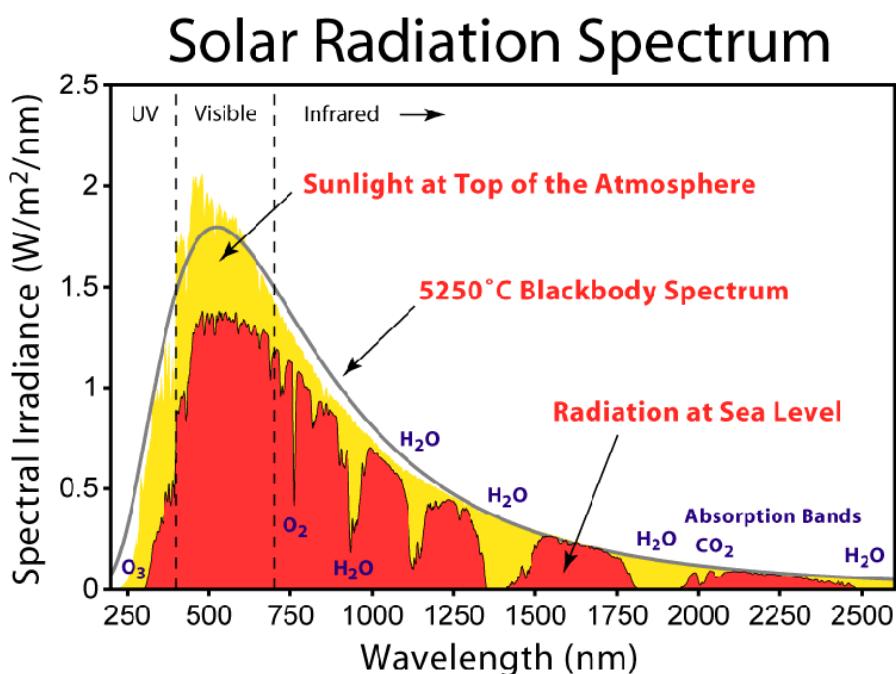


Figure 11.5: Solar radiation spectrum from 250 - 2500 nm. Irradiance measurements at the top of the atmosphere (yellow) and sea level (red) are depicted. The grey line represents the theoretical curve of a 5250 degree C blackbody spectrum [Rohde, 2008]. CC BY-SA 3.0.

The process of scattering is also affected by the properties with which the scattering EMR interacts. The angle at which EMR interacts with a surface, as well as the surface material, determine the properties of reflection. A reflector is described based on the properties of EMR that are reflected from it and range from specular to diffuse (Figure 11.6). Specular reflectance occurs when EMR is reflected in a single direction and can also be called anisotropic reflectance. A mirror is an example of a specular reflector. A diffuse, or Lambertian, reflector reflects EMR in all directions equally and can also be called an isotropic reflector. An example of a surface that reflects EMR isotropically is paper.

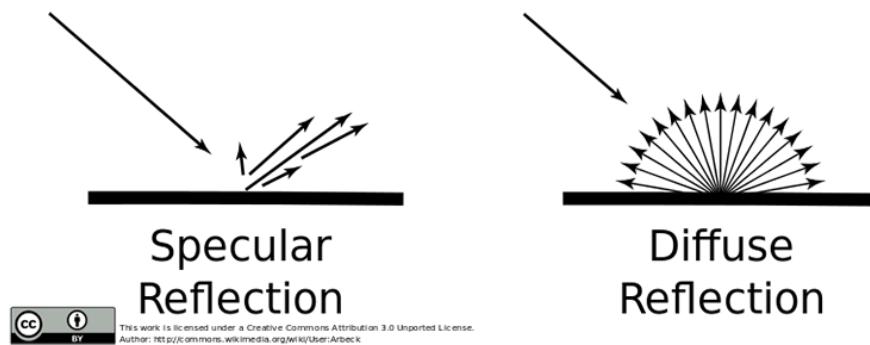


Figure 11.6: Depictions of specular and diffuse reflectors [Beck, 2012a]. CC BY-SA 3.0.

This is not to say that diffuse reflectors are perfect, however, as some issues remain related to the angle of incidence and the position of the sensor. For example, both back scattering and forward scattering affect the amount of radiation that reaches a sensor, depending on where the sensor is located. If a sensor is observing an object at the same angle as the incident radiation, the majority of reflected EMR will be from backscatter, or EMR that is scattered back towards its source. If the object being observed is perfectly specular, no EMR reflected off the object would be captured by the sensor. If the object is a diffuse or near-perfect diffuse reflector, then there is less of a concern with regards to capturing reflected EMR.

## 11.9 Foundations of Measurement

Now that we have discussed radiant energy and the concepts underpinning its interactions with other objects, we can begin to explore the measurements that our sensors record. One of the most important concepts to understand is that of the spectral signature, or spectra. A spectral signature refers to the amount of electromagnetic energy recorded across a defined section of the EMS. A nice example of a spectral signature is Figure 11.5, which presents the sun's radiation between 250 - 2500 nm in the units of solar irradiance ( $\text{W}/\text{m}^2/\text{nm}$ ). Similar graphs are common throughout remote sensing and can employ different units of measure.

The base measurements taken to generate spectral signatures is of an objects radiance. Acquiring radiance across a defined section of the EMS can be conducted by a variety of sensors and at different spatial scales, highlighting the practical advantages of evaluating surfaces using EMR. To fully capture and compare the objects being measured, however, it is often necessary to normalize radiance. The need for normalization stems mainly from the aforementioned issues of atmospheric effects, source and sensor location and sensor calibration. As with any normalization, the first step is to identify our minimum and maximum values.

There are two common reference measurements used to determine minimum and maximum radiance: dark and white reference. A dark reference is often taken by measuring the amount of energy recorded by a sensor when the input device is ignored. In theory, this would be the internal darkness of the machine and is considered to be the minimum radiance value in practice. The maximum radiance value is slight more challenging to determine as it requires a perfectly diffuse, flat white surface. A commonly used material is Spectralon, which has almost 100% reflectance between 400 - 1500 nm and greater than 95% reflectance over the entire optical region of the EMS (250 - 2500 nm). With both minimum and maximum values defined, it becomes possible to calculate normalized spectral values for a variety of properties across changing conditions.

## 11.10 Methods of Normalization

Upon calibrating an instrument to both 100% and 0% reflectance, it is possible to determine three normalized measurements of EMR: reflectance, transmittance and absorption. Each of these measurements provides useful information for understanding the interactions between EMR and the environment.

Reflectance refers to EMR that has interacted with and been reflected by a surface. It has emerged as a popular method of evaluating a variety of environmental properties, including land use change, plant health and plant diversity [Asner et al., 2011]. Another popular normalized measure of the interaction between photons and a surface is transmittance. A photon that is transmitted has passed through the surface with which it interacted and provides insight

regarding how much energy can reach other surfaces below. In a forestry context, this information can be particularly useful when determining the amount of radiation that reaches below the upper canopy (cite LAI, etc.). Absorptance is a third, related measurement that refers to the amount of energy absorbed by the cells within a surface and is roughly equal to the amount of energy not captured as reflectance or transmittance (Equation (11.7)).

$$\text{Absorptance} = 1 - \text{Reflectance} - \text{Transmittance} \quad (11.7)$$

Although relatively straight forward, these definitions allow us to start exploring a variety of remote sensing applications. In fact, most optical remote sensing techniques employ at least one of reflectance, transmittance and absorptance to examine the world. Before moving on to the next section, please review the work flow below highlighting what we have learned so far. In our next steps we will move from theory to application and begin to explore the factors that define the quality, and therefore capability, of remotely sensed data.

## 11.11 The Four Resolutions

One of the first considerations any user must make regarding remotely sensed data is its quality. For most scientific research, good quality data needs to contain information that is relevant to the scale and time period of the study, meaning that sensors are often designed for a specific application. The manner in which we describe a sensor is through four key resolutions.

## 11.12 Spatial Resolution

Although each resolution is important, spatial resolution holds a key position when determine the usefulness of a dataset as it determines the scale at which information is collected. When a sensor collects information it does so in a single area. That area could be the size of a single tree or a single city, but all the EMR measured by the sensor will be an average of that area. Generally, this area is referred to as a picture element, or **pixel**. A pixel is the smallest addressable digital element and basic unit of remotely sensed data. When multiple pixels are collected in adjacent areas, perhaps using an instrument with multiple sensors on it, the output is called an image. In short, an image is a collection of pixels, which represent mean values of reflected or transmitted radiation over some area. Spatial resolution, then, is the ground area represented by a pixel.

There are a variety of factors that affect spatial resolution, or the size of a pixel. One important factor is the sensor's field of view (FOV). A field of view refers to the observable area of a sensor and is defined by two things: the angle of the FOV and the sensors distance from it's target. Changes in these two factors result in an increase or decrease in the amount of area captured by a sensor and therefore a change in pixel size. Pixels that cover larger areas are considered to

have lower spatial resolution, while a relatively smaller pixel is considered high spatial resolution. When a sensor is in motion, collecting multiple pixels across space and time, the term instantaneous field of view (IFOV) is used to describe the FOV at the time each pixel was collected. We will learn more about the challenges of collecting data over space and time in Chapter 12 and Chapter 15.

It is a trade off when determining if a certain spatial resolution is useful for a given application, the pixel size needs to be fine enough to observe features of interest, but large enough to be stored and processed in a reasonable manner. It must also cover the entire study area, which can vary significantly depending on the research objectives and the image's pixel size. Each of these considerations will direct the data user to a specific sensor. From here, the user can begin to consider the remaining three resolutions.

## 11.13 Temporal Resolution

Much like spatial resolution deals with the space that a sensor observes, temporal resolution refers to the time interval between successive observations of a given space. Temporal resolution can span seconds or years and is requirement when investigating change. An acceptable temporal resolution is defined inherently by the nature of the study. For example, a study monitoring the annual urban expansion of Vancouver, B.C. would have a temporal resolution of 1 year. Other common temporal resolutions include hourly, for applications monitoring weather, and daily, which is often used when monitoring plant phenology.

The interval between successive observations is often called *revisit time* and is determined by the type of platform on which a sensor is deployed. Sensors that remain stationary can have revisit times measured in seconds, while sensors moving faster or slower than their target area can take multiple days before they observe a location again. Interestingly, sensors have been designed to be both stationary and mobile at a variety of spatial scales. Both airborne and spaceborne sensors have the capacity to remain stationary over a single location or travel faster than the Earth rotates and capture imagery over most of the Earth's surface in a matter of days.

## 11.14 Spectral Resolution

Earlier in this chapter the concepts and theories surrounding EMR were presented. These theories related directly to the concept of spectral resolution, which refers to the number and dimension of specific EMR wavelengths that a remote sensing instrument can measure. Due to the large range of the EMS and properties of EMR, the term spectral resolution is often used to refer to any single component of its definition. In scientific literature, it is not uncommon to find "spectral resolution" referring to:

- the number of spectral bands (discrete regions of the EMS) that are sensed

as a single unit.

- the location of these units, or groups of bands, along the EMS.
- the number of individual bands within each unit. Also called bandwidth.

Each of these components plays a role in describing the spectral resolution of a pixel and enables users to identify appropriate sensors for their application. It is also important to consider the laws associated with the energy of EMR. Recall that shorter wavelengths have more energy, which makes it easier to detect. The implications of this is that longer wavelengths require larger bandwidths for the sensor to observe them. An easy visualization of this concept involves selecting two wavelengths along the EMS. If we select the first wavelength at 0.4 nm and the second wavelength at 0.8 nm, we can use Equation (11.1) to demonstrate that the first wavelength has twice as much energy as the second. This is an important theory to note as the consequences of a decrease in energy is a decrease in spatial resolution (a larger number of bands need to be combined to collect enough information).

A common method for visualizing the spectral resolution of a sensor is to place each band along the EMS according to its associated bandwidth and wavelengths (Figure 11.7). This allows users to determine which sensor best captures the information they are interested in studying. For some applications, such as land cover, it is acceptable to use sensors with relatively wide bands collecting information in a small number of strategic locations along the EMS. For other applications, spectral information may need to be more detailed and capture information using thin, adjacent bands spanning a large region of the EMS. These specifications will be discussed in greater detail in Chapter 12, so for now we'll focus on how spectral information can be useful.

The collection of spectral data across more than one band allows the creation of a spectral curve, or spectral signature. Spectral signature are the cornerstone of many remote sensing applications and highlight many properties of the surface from which they were collected. The creation of a spectral signature is quite simple and can be depicted in two dimensions (Figure 11.8). Essentially, the observed value of each band is connected to the observed value of each adjacent band in 2D space. When all bands are connected, a spectral signature is born. As we will see in a later case study, spectral signatures can provide a plethora of relevant information relating to the composition of an object.

## 11.15 Radiometric Resolution

In short, radiometric resolution is the quantification of a sensor's ability to detect differences in energy. Photons enter a sensor through a filter that only permits specific wavelengths. The energy of the photon is recorded as a digital number (DN) and the digital number is assigned to a pixel.

A simple visualization of radiometric resolution would be to think of three colors:

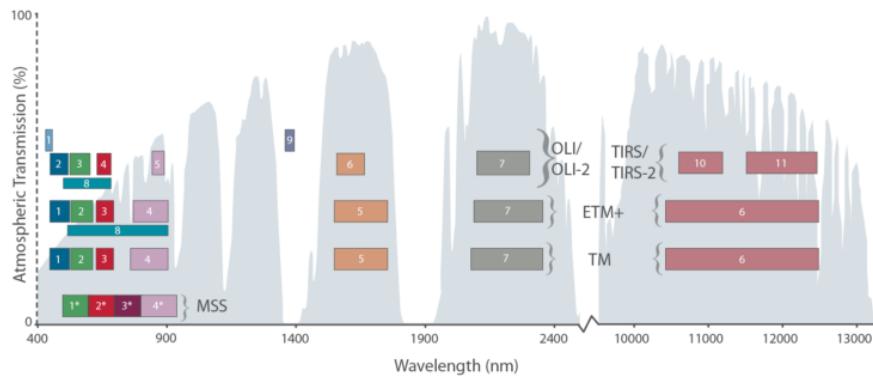


Figure 11.7: Locations of bands for various sensors deployed by NASA on one of more Landsat misison. Landsat 1-5 had the Multispectral Scanner System (MSS), while the Thematic Mapper (TM) was aboard Landsat 4-5. The Enriched TM Plus (ETM+) had 8 bands and was aboard Landsat 7. Grey distributions in the background represent the atmospheric transmission values for a mid-latitude, hazy, summer atmosphere [NASA, a].

red, green and blue. A sensor detecting energy in the ranges of these three wavelengths would contain three separate detectors. Each detector is specialized to record energy in a single, unique range, say red ( $\sim 700$  nm). The amount of energy that is recorded while observing an area is stored in a pixel as a DN, with the lowest DN number representing zero sensor detected energy in this wavelength range and the highest DN representing maximum sensor detected energy.

The radiometric resolution of a detector, then, is the number of discernible values, or segments, present between zero and maximum DN values. These segments are usually referred to as bits and can be mathematically represented as an exponent of 2 (Equation (11.8)). The number of bits in that a detector can resolve may also be called the *colour depth* of the image. Figure 11.9 clearly presents the increase in detail provided by additional bits, but recall that increasing any resolution generally increases storage and processing time. As such, it is important to select an appropriate radiometric resolution based on the needs of your study.

$$8\text{bits} = 2^8 = 256\text{levels} \quad (11.8)$$

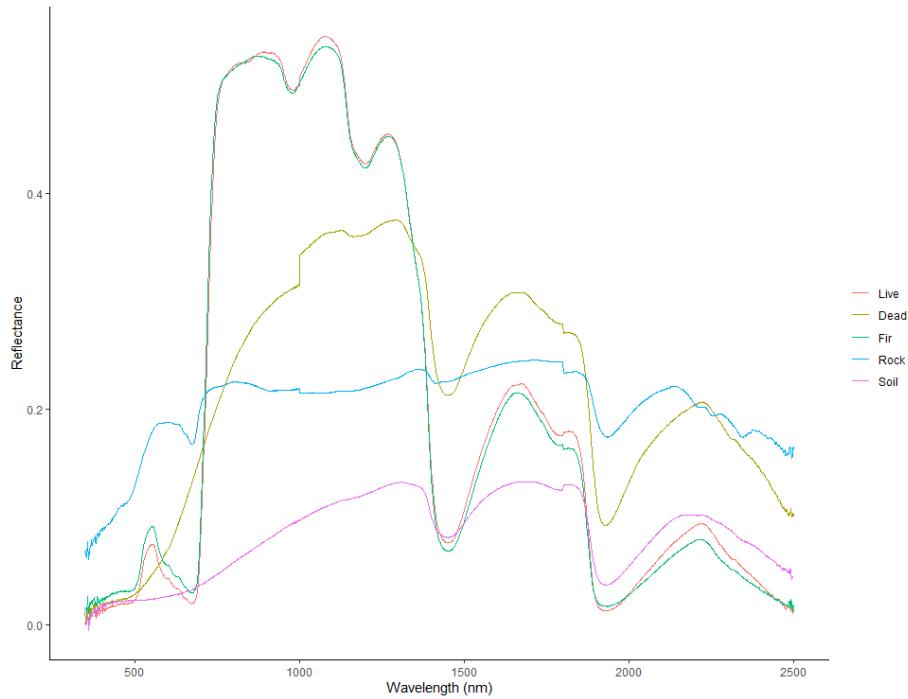


Figure 11.8: Five spectral signatures of various living and non-living samples collected using an ASD FieldSpec3 Imaging Spectroradiometer. Live = live broadleaf, Dead = dead broadleaf, Fir = Douglas-fir needles on branch, Rock = rock, soil = dry humus soil.

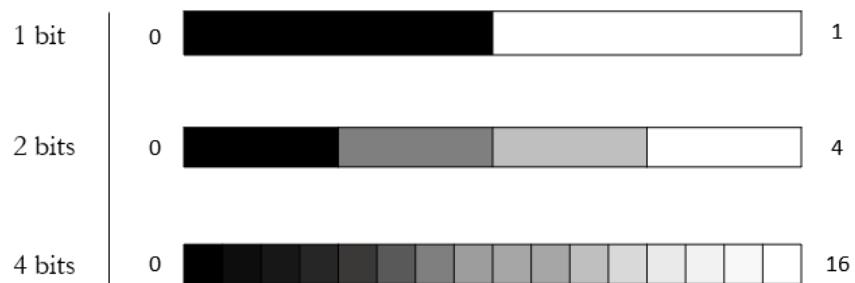


Figure 11.9: Discernable values of sensors with 1, 2 and 4 bit radiometric resolutions. Created by Paul Hacker CC BY-SA 4.0.

## 11.16 Key Applications

So far in this chapter we have covered the theories and concepts that justify the use of EMR for remote sensing. With these fundamentals in mind, we can begin discussing some common applications of remote sensing. For the purposes of this book, we will focus on studies related to environmental management.

The use of optical remote sensing (400 - 2500 nm) to analyze the environment has become popular over the past half century. Sensor development, improved deployability and decreasing costs have enabled many researchers to use selected sections of the EMS to monitor everything from the chlorophyll content of a single leaf [Curran, 1989] to global forest cover [Hansen et al., 2013].

Large-scale research projects focused at national or international levels have perhaps benefit the most from improved sensor deployment. Since the 1960s, a variety of satellites have been launched with the sole purpose of observing the Earth. Landsat, created by NASA and is now run by the United States Geological Survey (USGS), is a suite of satellites designed specifically for this purpose. By the end of 2021, a total of nine Landsat missions will have been launched, eight of which have successfully reached orbit and provided imagery in at least 5 broad spectral bands at  $30\text{ m}^{-2}$  spatial resolution. This information has been used to monitor of land cover change, ecosystem services and a variety of other environmentally relevant metrics [Deel et al., 2012]. The case study at the end of this section highlights a particularly novel approach to optical remote sensing that has become a popular methodology to evaluate plant health and biodiversity [Ustin and Schaepman, 2009, Wang et al., 2018].

As far as Canada's contributions to remote sensing sensors rank, RADARSat is among the most important [Raney et al., 1991]. This satellite was launched in 1995 and has grown to a constellation of three space-borne synthetic aperture radar (SAR) sensors that feature variable resolution. As an active sensor, RADARSat produces and measures EMR with wavelengths between 7.5 - 15 cm and is capable of penetrating clouds and smoke. These characteristics make RADARSat ideal for applications such as ecosystem monitoring, maritime surveillance and disaster management. It's active nature also enables RADARSat to record observations at night. Chapter 12 will discuss radar in more detail.

Another active remote sensing technique that has become popular is light detection and ranging (LiDAR), which can also be called airborne laser scanning (ALS). LiDAR is particularly useful in evaluating structural components of the environment, such as forest canopies and elevation [Coops et al., 2007]. More details regarding the theories and applications of LIDAR will be presented in Chapter 15.

### 11.17 Case Study: Optical Remote Sensing to Evaluate Land Cover in Canada

The capacity to evaluate change in an environment relies on the ability to capture information at useful resolutions. Land cover, which is simply the combination of abiotic and biotic components existing in a defined area, is a useful metric for evaluating changes across an ecosystem and land cover analyses have been deployed at a variety of scales. Land cover is also an essential climate variable and is considered an important metric that can aid in the prediction of future climate regimes. These studies can be limited, however, by the resolutions at which information is collected. Hermosilla et al. provide an interesting method of land cover assessment that utilizes Landsat to derive land cover and land cover change across Canada over a 29 year span [Hermosilla et al., 2018].

Before exploring the methods of this study, it is important to highlight the qualities of the Landsat program that enabled it. Launched in 1972, Landsat was NASA's first satellite designated for evaluating Earth's surface. Although a variety of technological upgrades have been introduced since the 1972 launch, the majority of Landsat sensors have a spatial resolution of  $30m^2$ . Considering the spatial coverage of this mission includes the majority of our planet, this spatial resolution is quite useful.

Another important quality of Landsat is that every satellite carries a multi-spectral sensor with at least four bands, which enables the differentiation between multiple land cover classes, including vegetation and urban building. More information regarding the changes in spectral resolution between Landsat satellites can be found in Figure 11.7. The temporal resolution of Landsat is the third high-quality characteristic of this satellite mission that enables the accurate evaluation of land cover. With a return interval of 16 days, Landsat provides information for the same location on Earth almost twice per month. This is particularly important for a country as large as Canada, which contains 12 ecozones dominated by forest, and enables the selection of a Best Available Pixel (BAP), which is selected based on a variety of criteria that quantifies a pixels quality (Figure 11.10). You can learn more about Landsat BAP in the article by Thompson et al. [Thompson et al., 2015].

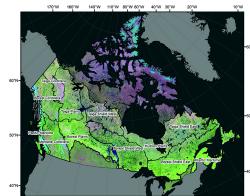


Figure 11.10: False colour composite (bands: shortwave infrared, near infrared and red) for 2010. Each pixel in this image was selected based on its classification as a Best Available Pixel (BAP) [Hermosilla et al., 2018].

For their analysis, the authors Hermosilla et al. used imagery captured from Landsat 5 onward with the Thematic Mapper, which means their data contained seven unique spectral bands spanning segments between 0.45 nm to 2.35 nm. The ability to observe reflectance values across a relatively wide range of the EMS enabled land cover analysis across all forested eco-regions in Canada. After applying a masking model, the authors devised a framework to generate annual land cover maps for all of Canada's forested areas, which were presented at the beginning of this chapter (Figure 11.2). From these annual composites, it was also possible for the authors to identify land cover change over time and quantify the effects of specific disturbance types, such as harvesting and fire (Figure 11.11). These comprehensive maps demonstrate the capability of well-designed remote sensing technologies to obtain scientifically relevant information and highlight the potential of remote sensing in environmental assessment [Hermosilla et al., 2018].

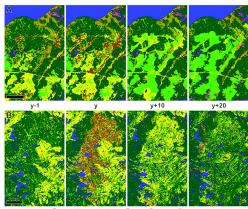


Figure 11.11: Temporal sequences depicting landcover before and after (a) harvesting and (b) fire. The year of change is represented by 'Y' [Hermosilla et al., 2018].

## 11.18 Summary

In this chapter we have covered a variety of physical theories that support the use of electromagnetic **radiation** for the remote analysis of objects. These ideas provide the fundamental knowledge needed to employ remote sensing technologies. From sensor selection to data processing, your ability to perform remote sensing science will rely almost exclusively on your capacity to comprehend the relationships between the objects of interest and the physical properties that comprise them.

In the next chapter you will enter the world of sensors and explore a vast amount of technologies that collect information remotely. This journey will include active and passive instruments, as well as examples of scientific studies that have successfully employed the data they collect. Much like a carpenter selecting the correct tool for a specific task, you will need apply your knowledge of the fundamentals to identify the best sensor for your research.

### **11.19 Reflection Questions**

1. What section of the electromagnetic spectrum is adjacent to, but shorter in wavelength than, what the human eye can see?
2. What is the general term used to describe the process that stops large amounts of electromagnetic radiation from reaching the Earth's surface. Name the three types?
3. List and describe the four resolutions.
4. Why is using the Best Available Pixel important for land cover mapping?

## Chapter 12

# Remote Sensing Systems

Written by Claire Armour and Paul Pickell

You probably know that you are using your very own organic remote sensing system to read this sentence. Our eyes take in information from the world around us by detecting changes in light and relaying that information through the optic nerve into our brains, where we make sense of what we are seeing. As you learned in Chapter 11, this is what constitutes remote sensing - gathering information (“sensing”) without directly measuring or interacting with that information (“remote”). Whereas our eyes are limited to the visible light portion of the electromagnetic spectrum and by the location of our bodies, remote sensing systems use powerful sensors and flight-equipped platforms to paint a broader and deeper picture of the world around us. The picture in figure 12.1 is an example of the beautiful imagery we can capture from space, taken from the GOES-1 satellite.

Remote sensing systems range in size and complexity from a handheld camera to the Hubble telescope and capture images of areas ranging from a few meters to several kilometers in size. Though devices such as microscopes, X-ray machines, and handheld radios are technically remote sensing systems, the field of remote sensing typically refers to observing Earth on a small spatial scale (1:100 to 1:250,000).

### NOTE

Remember, in spatial scale, “small” means a big picture. If you want a refresher on how to read and understand map scales, check out section 2.5 in Chapter 2.

The range of uses for remote sensing platforms are dazzling in number, allowing us to monitor severe weather events, ocean currents, land cover change, natural disturbances, forest health, surface temperature, cloud cover, urban development, and so much more with high precision and accuracy. In this chapter, we

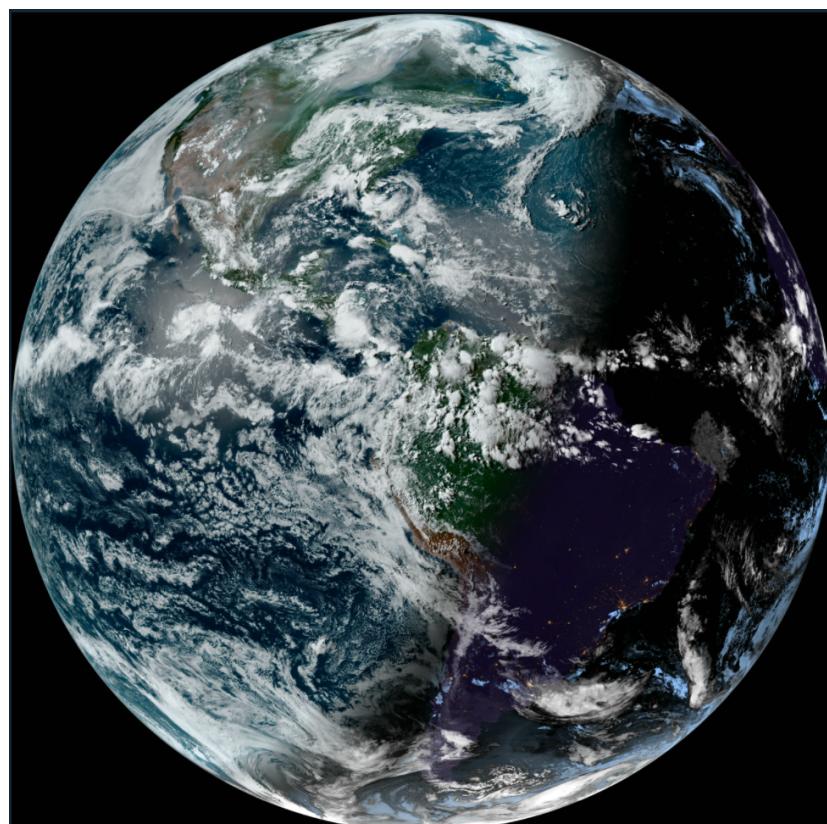


Figure 12.1: North and South America as seen from the NASA GOES-1 satellite [NASA, b]. Captured from KeepTrack.space. Copyright (C) 2007 Free Software Foundation, Inc.

will break down the how and where of remote sensing systems and discover a few different systems used for Earth observation today.

### Learning Objectives

1. Break down remote sensing technology into its basic components
2. Understand how different settings and parameters impact remote sensing system outcomes
3. Review the key remote sensing systems used in Canada and around the world for environmental management

### Key Terms

Absorption, Aerial, Along-Track, Atmospheric Window, Biconcave, Biconvex, Convex, Concave, Cross-Track, Curvature, Field of View (FOV) Focus, Geosynchronous Equitorial Orbit (GEO), Hyperspectral, Instantaneous Field of View (IFOV), Low Earth Orbit (LEO), Medium Earth Orbit (MEO), Multispectral, Nadir, Near-Polar Orbit, Oblique, Orbit, Panchromatic, Pitch, Push broom Scanner, Radiometric, Radius of Curvature, Reflection, Refraction, Resolution, Roll, Spectral, Sun-Synchronous Orbit, Thermal, Whisk Broom Scanner, Yaw, Zenith

## 12.1 Optical System Basics

Remote sensing systems contain a number of common components and operate using similar principles despite their differences in capabilities. In the subsequent sections, we will discover the technical specifications of remote sensing systems that allow them to “see”.

### 12.1.1 Lenses

Picture the view from a window onto a busy street on a rainy day: you have cars driving by with headlights, traffic lights reflecting off a wet road, raindrops pouring down the windowpane and distorting the view, and hundreds of people and objects on the street scattering light beams in every possible direction from a huge range of distances. In order for us to take in any of this, these light beams need to reach the retina, the photosensitive surface at the very back of the eye. How do our relatively tiny eyeballs take in all that disparate light and produce crystal clear images for our brains? By using one of the most basic components of any optical system: a lens. A lens is a specially shaped piece of transparent material that, when light passes through it, changes the shape and direction of light waves in a desired way.

The property of transparent mediums to change the direction of light beams is called **refraction**, or transmittance. This is why objects in moving water look misshapen. The arrangement of molecules within a medium disrupts both

the direction and speed of the photons – the measurement of this disruption is called the refractive index.

The lens at the front of your eyeball changes refracts light beams from varying distances precisely onto your retina. The optical systems on remote sensing platforms are the same: they use a specially designed lens to focus light beams at the desired distance to onto their own recording medium. Below in figure 12.2 is a simple visualization of how optical systems focus light onto a desired point to produce an in-focus image.

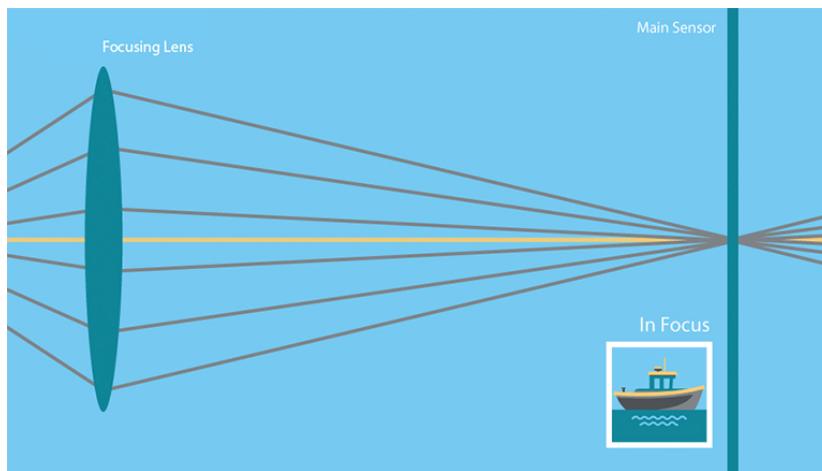


Figure 12.2: Focusing lens [Vorenkamp, 2015]. Copyright Todd Vorenkamp. Used with permission. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengis.ca/remote-sensing-systems.html#fig:12-focus-example-humaneye>.

Now, picture another scene: you are scrolling through social media and you see a beautiful photo of Mt Assiniboine taken by your friend, a professional photographer based in the Canadian Rockies. You think to yourself, “Wow, that peak looks ENORMOUS! I want to visit there and see it for myself.” So, you ask your friend exactly where they went, drive to Banff National Park, hike to the very same spot, and squint skywards. Hmm...though the mountain is still imposing, it is certainly not towering over you at close range as it was in the photo. You also notice that there are several surrounding peaks that you couldn’t see before. Your friend’s picture was crystal-clear, and you have 20/20 vision, so you know its not an issue with focusing properly. Are you being deceived?

Actually, yes, you are – by both your eyes *and* the camera your friend used. We like to think that our eyes show us the world as it truly is and that everything else is a facsimile, but in truth, all optical systems alter the scenes around us to show us what we need to see. From an evolution standpoint, you can see why

clear resolution of close-range objects would be of vital importance for humans – think distinguishing edible plants from poisonous ones, hunting prey, reading facial expressions, etc. We can make out human-sized objects up to a distance of three kilometres in good lighting [Wolchover, 2012], but if you are interested in seeing something far away, such as a mountainside or a celestial body, you'll have to trade in your natural close-range viewing abilities for a system specialized for distant details – e.g., binoculars or a telescope. The distance at which objects can be resolved and how they appear in an image lies with the lens. Read on below to learn about how different lens designs influence the appearance of a scene or object, and keep in mind how these designs may be used in various earth observation applications.

Most, if not all, lenses on optical systems for remote sensing are **spherical lenses**, called that because each side of the lens is spherical in shape, similar to a bowl. A **convex** optical surface curves outward from the lens centre, whereas a **concave** optical surface curves inward toward the lens centre. Though not spherical, a planar or flat optical surface may be used as well. A spherical lens is formed by joining two optical surfaces – concave, convex, and/or planar – back-to-back. A **biconvex** or positive lens is two convex surfaces, and a **biconcave** or negative lens is – you guessed it – two concave surfaces. The **radius of curvature** is the measure of how much an optical surface “bulges” or “caves”. If you imagine tracing the edge of the surface in an arc and continuing the curve all the way around in a circle, the radius of this imagined circle would be the radius of curvature. Biconvex and biconcave lenses can be “equiconvex”, meaning they have the same spherical curvature on each side, but may also have uneven curvatures. The lens in the human eye is an example of a lens with uneven curvatures – our radius of curvature is higher at the front. Figures 12.3 and 12.4 demonstrate how the radius of curvature is measured for both concave and convex optical surfaces.

### Your Turn!

What is the radius of curvature for a perfectly flat lens? See answer at end of chapter.

#### 12.1.2 Focal Length

Now that we know a little bit about what lenses look like, let us turn to consider how an image is projected onto the recording medium. As you might expect, different combinations of optical surfaces and radii of curvature will behave in different ways. Remember that for an image to be in-focus, we need to ensure the light beams are landing precisely on the recording medium or screen.

Convex optical surfaces cause light beams to *converge*, or focus, to a point behind the lens.

Concave lenses cause light beams to *diverge*, or spread out, resulting

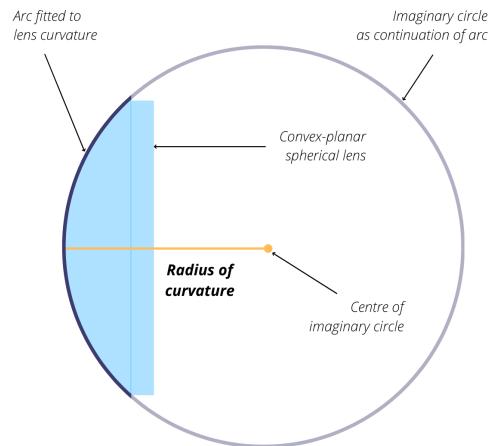


Figure 12.3: Measuring the radius of curvature for a convex optical surface.  
Claire Armour, CC BY 4.0.

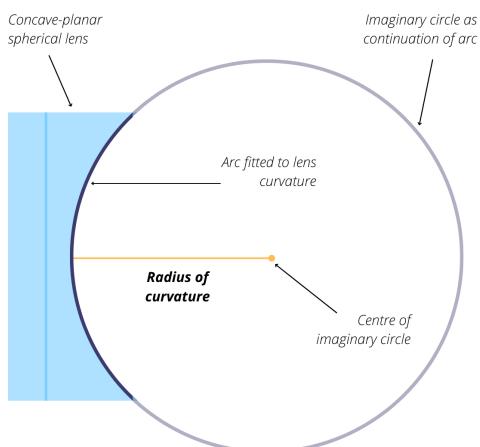


Figure 12.4: Measuring the radius of curvature for a concave optical surface.  
Claire Armour, CC BY 4.0.

in the light appearing to converge (focus) to a point in front of the lens.

The point where the light converges or appears to converge is called the *focal point*, and the distance between the focal point and the centre of the lens is called the *focal length*. For a converging lens, the focal length is positive; for a diverging lens, it is negative. Figures 12.5 and 12.6 illustrate the behaviour of light when travelling through a biconvex and biconcave lens. See the paragraph below the diagrams for variable labels.

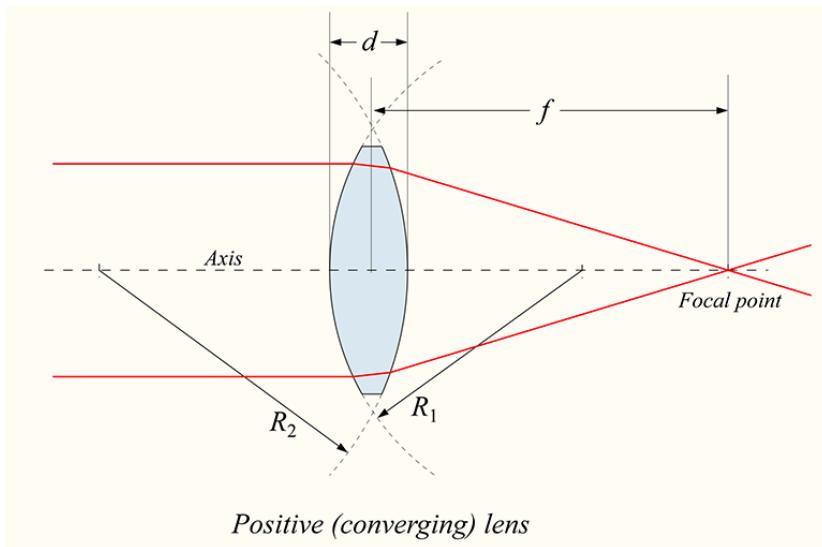


Figure 12.5: Measurements in a biconvex lens. [DrBob, 2006b], CC BY 3.0 Unported.)

The optical power of a lens – the degree to which it can converge or diverge light – is the reciprocal of focal length. Essentially, a “powerful” lens will be able to refract light beams at sharper angles from the horizontal, causing them to converge or appear to converge closer to the lens, i.e., at a smaller focal length. The **Lensmaker's Equation** (Equation 1) allows us to calculate the focal length ( $f$ ) and/or optical power ( $\frac{1}{f}$ ) as a function of the radii of curvature ( $R$ ), the thickness of the lens between the optical surfaces ( $d$ ), and the refractive index of the lens material ( $n$ ). Note that  $R_1$  is the front surface - the side of the lens closest to the origin of the light - and  $R_2$  is the back surface.

Equation 1:

$$\frac{1}{f} = (n - 1) \left[ \frac{1}{R_1} - \frac{1}{R_2} + \frac{(n - 1)d}{nR_1 R_2} \right]$$

We know how lenses impact focal length, but how does focal length impact a photo? Let us return to our scenario in Banff National Park where we have two

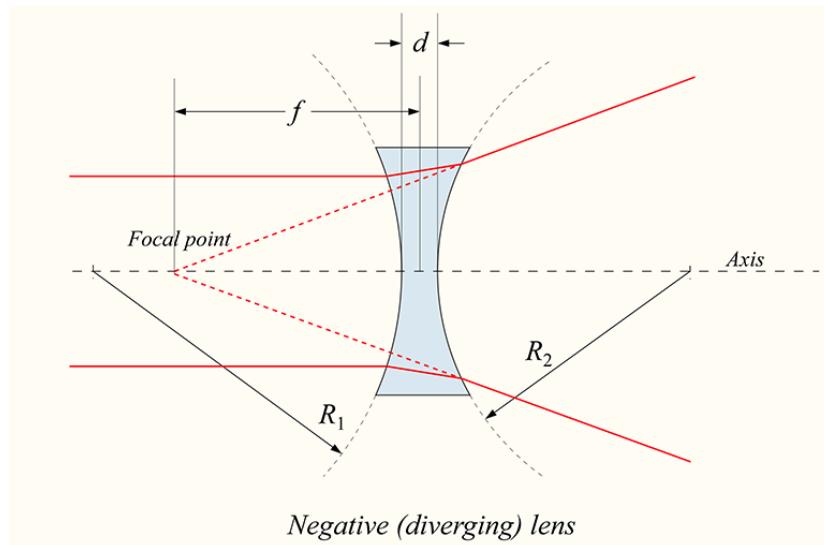


Figure 12.6: Measurements in a biconcave lens. [DrBob, 2006a], CC BY 3.0 Unported.

mismatched images of the same mountain. You've done some investigating and found that the camera your friend used is very large (and expensive). The lens at the front is quite far from the recording medium in the body of the camera – many times the distance between your own eye lens and recording medium, the retina. This difference in focal length is the cause of the differing images. A low optical power lens with a long focal length will have high magnification, causing distant objects to appear larger and narrowing the field of view (see section 12.something). A high optical power lens with a short focal length, such as your eye, will have low magnification and a larger field of view by comparison. Mystery solved!

### Your Turn!

When you return from your trip, your friend decides to test you on your new skills and shows you these additional photos they took of Mt Assiniboine in nearly identical spots on the same day. They ask you which photo was taken with a longer camera lens. How can you know? *Try this:* from the peak of Mt Assiniboine (the very big one), draw a line straight downwards or cover half of the photo with a piece of paper, and then do the same for the other photo. Does the line or paper edge intersect at the same points of the foreground in each photo? Can you see the same parts of the mountains in the foreground? Use the rock and snow patterns for reference. If the cameras were the same focal length, even with different cropping and lighting as seen here, the answers should both be yes. Can you tell which photo was taken with a 67mm lens and

which was taken with a 105mm lens? See the answers at the end of the chapter.



Figure 12.7: Mt Assiniboine, image one [Maguire, a], CC BY 4.0.)



Figure 12.8: Mt Assiniboine, image two [Maguire, b], CC BY 4.0.

### 12.1.3 Sensors

The **sensor** is the subsystem that is responsible for digitally recording the intensity of electromagnetic radiation. Sensors are engineered in different ways to suit the type of electromagnetic radiation that needs to be recorded. However, most sensors will be comprised by three primary elements: a filter, an array of detectors, and an analog-to-digital converter. The **filter** is responsible for ensuring that only the desired wavelength of electromagnetic radiation enters the sensor. For example, in order to image only the near infrared part of the spectrum, a

filter is needed to block out photons of all other wavelengths from entering the sensor. Once the desired wavelength of electromagnetic radiation is in the sensor, then the photons fall onto **detectors**, which are responsible for recording the electromagnetic radiation at a specific location in the array. Finally, the **analog-to-digital converter** is responsible for converting the photon energy into a measurable electrical charge that eventually becomes the digital number in the image for a given pixel. In summary, the photons enter the telescope or camera lens, are focused onto a plane through a lens, then filtered by wavelength, then they fall onto individual detectors before being converted to digital signals that represent numbers in a raster image. This whole process is illustrated in Figure 12.9 below.

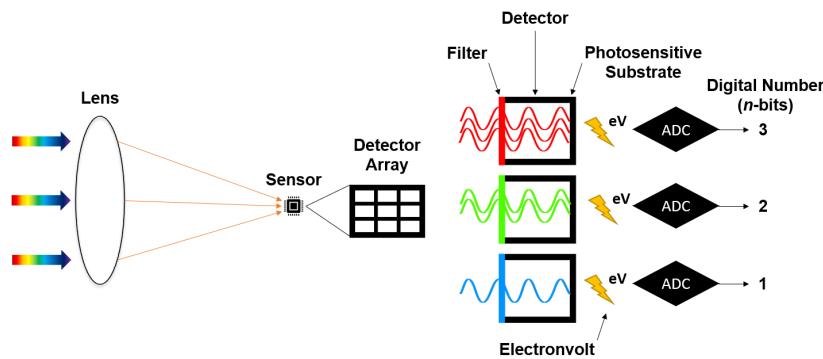


Figure 12.9: Electromagnetic radiation enters the lens, where it is refracted and focused onto a surface containing the digital sensor. An array of detectors are arranged on the digital sensor that represent different pixel locations in the output raster image. Filters are used to ensure only specific wavelengths are recorded by each detector. Energy from photons is converted to electrical charges and then converted to digital numbers by the analog-to-digital converter. Pickell, CC-BY-SA-4.0.

#### 12.1.4 Field of View

When you look at something, there are likely other objects you can see above, below, and beside it through your peripheral vision. Remote sensing systems have an analogous **field of view (FOV)** that describes the angular range of observation. By contrast, the **instantaneous field of view (IFOV)** describes the angular range of what an individual detector can observe. In other words, the FOV tells us what the remote sensing system is capable of seeing through its entire range of motion and the IFOV tells us what the a single detector can see in a given moment. Both of these measures are important for describing the quality of the imagery that is collected, both in terms of how large of an area can be imaged as well as the spatial resolution of the imagery.

Building remote sensing systems can be costly, so it is in our best interest to

have them see as much as possible with the least expenditure of effort. We can maximize the FOV of a remote sensing system by giving it the freedom to “look around”. Humans have three degrees of motion that allow us to change our FOV: scanning with our eyes, swiveling our heads, and shifting the position of our bodies. Remote sensing systems can have three analogous degrees of motion: the motion of the lens elements (eyes - analogous to focus or zoom), the motion of the camera (head - analogous to scanning), and the motion of the platform (body - analogous to direction of travel).

Remote sensing systems typically have zero, one, or two degrees of motion. Rarely do remote sensing systems have all three. It is usually not necessary to have so much range of motion and more moving parts means a higher possibility of malfunction, which can be a real headache when the defunct system is orbiting 700 km above the Earth’s surface. These combinations of degrees of motion give rise to the three primary types of scanners. **Push broom** scanners are a type of scanner that have detectors arranged in a single-file line and take advantage of the forward movement of the remote sensing platform, known as **along-track** scanning, to build images line-by-line (Figure 12.10). In other words, push broom scanners have two degrees of motion: the platform and the lenses.

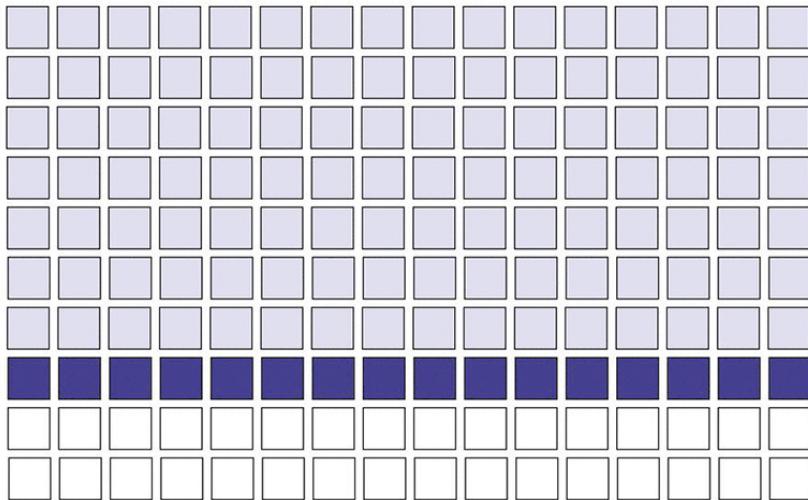


Figure 12.10: Visualization of how a push broom scanner captures imagery. The dark purple squares represent the subset of the area seen by the scanner at any given time and the lighter purple squares show previously scanned areas. Armour, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/remote-sensing-systems.html#fig:12-push-broom>.

By contrast, **whisk broom** scanners have an array of detectors that are mechanically moved from side-to-side, known as **cross-track** scanning because the

image is produced by scanning across the track of the remote sensing platform direction of motion ((Figure 12.11)). Thus, whisk broom scanners have three degrees of motion: the platform, the lenses, and the camera. The last type of scanner is known as a **staring array**, so-called because the sensors are arranged in a rectangular array that are pointed at the surface or object to be imaged. In this way, the image is built all at once as the light is focused onto the focal plane. Most consumer cameras use staring arrays, which usually have two or fewer degrees of motion, depending on the application.

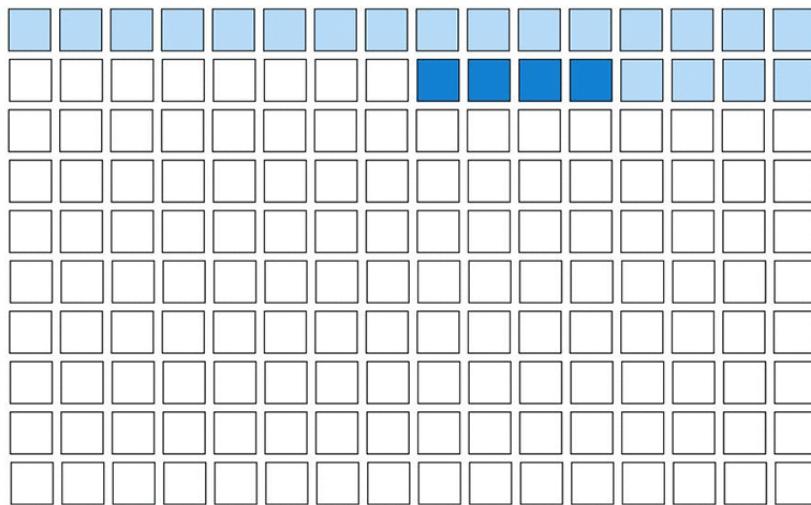


Figure 12.11: Visualization of how a whisk broom scanner captures imagery. The dark blue squares represent the subset of the area seen by the scanner at any given time and the lighter blue squares show previously scanned areas. The size of the subset may change between a single pixel (one square) or a spotlight (multiple squares) but the motion remains the same. Armour, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/remote-sensing-systems.html#fig:12-whisk-broom>.

In remote sensing systems, the FOV is usually expressed as an angle with the following equation:

$$FOV = 2 \times \theta + \beta$$

where  $\theta$  is the scan angle and  $\beta$  is the IFOV. The **scan angle** describes the physical limits of the system to mechanically turn from side-to-side (e.g., whisk broom sensor) or the physical limits of the incoming light to be refracted by the lens onto the focal plane (e.g., push broom sensor). We can also derive the distance on the vertical datum or ground, known as the **swath width (W)**, if we know the altitude or height that the sensor is at:

$$W = 2 \times H \times \tan(\theta + \beta/2)$$

where  $H$  is the height of the sensor above the vertical datum or ground. Figure 12.12 illustrates how height, IFOV, and scan angle are related.

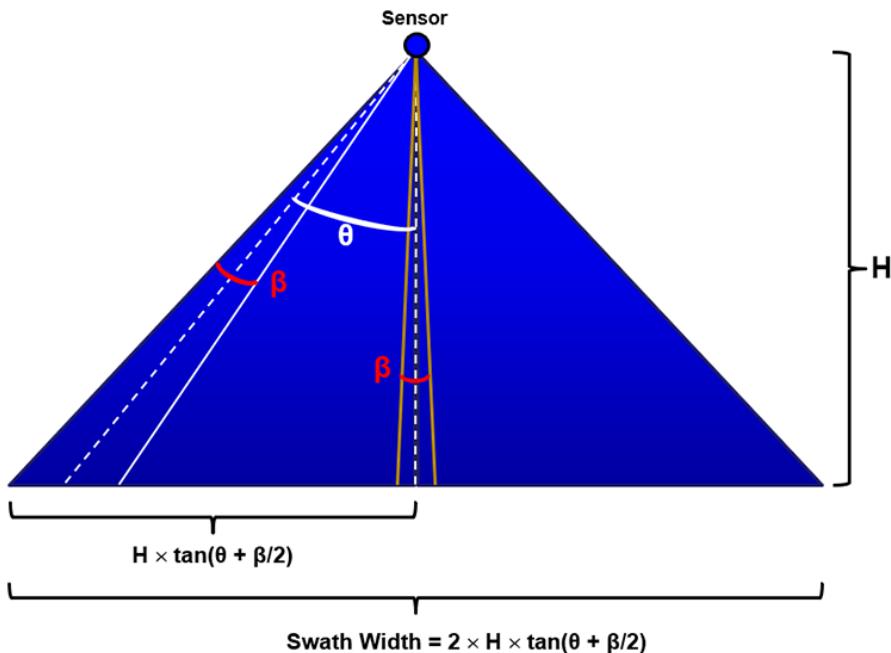


Figure 12.12: Swath width. Pickell, CC-BY-SA-4.0.

All of these parameters (scan angle, IFOV, and height) are incredibly important for how much of the ground a sensor is actually observing within a given pixel. In fact, pixels that are imaged at the edge of the focal plane will necessarily represent larger areas on the ground. Figure 12.13 illustrates how a pixel at nadir  $P_n$  will represent a ground distance equal to  $H \times \tan(\beta)$  while a pixel at the extreme of the scan angle will represent a ground distance equal to  $H \times \tan(\theta + \beta/2) - H \times \tan(\theta - \beta/2)$ . This is referred to as the “bow-tie” effect because pixel ground distance becomes elongated in both the cross-track and along-track dimensions (Figure 12.14) the farther you move away from nadir. The bow-tie effect is most evident when the scan angle exceeds 19°.

For example, the Visible Infrared Imaging Radiometer Suite (VIIRS) is a sensor aboard two weather satellites that orbit at an altitude of 829 km, have a scan angle of 56.28°, and a cross-track IFOV of 0.79°. With these parameters, the pixel cross-track ground distance at nadir is:

$$P_n = 829\text{km} \times \tan(0.79) = 11.43\text{km}$$

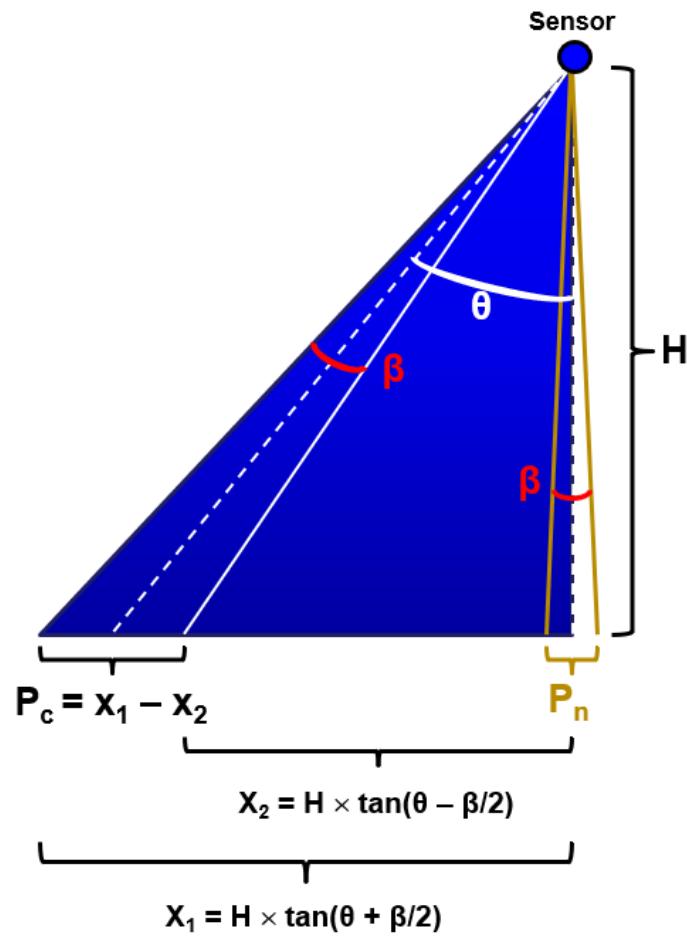


Figure 12.13: Scan angle. Pickell, CC-BY-SA-4.0.

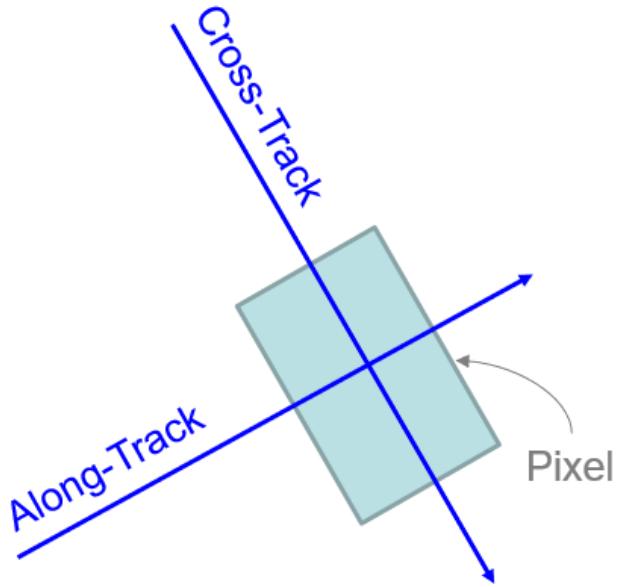


Figure 12.14: Pixel dimensions shown as a function of cross-track and along-track travel of the remote sensing platform. Pickell, CC-BY-SA-4.0.

and the pixel cross-track ground distance at the extreme of the scan angle is:

$$P_c = 829 \text{ km} \times \tan(56.28 + \frac{0.79}{2}) - 829 \text{ km} \times \tan(56.28 - \frac{0.79}{2}) = 37.09 \text{ km}$$

The cross-track ground distance of a pixel at the edge of a VIIRS image may be more than three times the cross-track ground distance at nadir!

## 12.2 Perspectives

All sighted creatures that we know of - save those from the water-dwelling genus *Copepoda* (<https://askdruniverse.wsu.edu/2016/05/31/are-there-creatures-on-earth-with-one-eye/>) - have two or more eyes. As the eyes are at different locations in space, each eye perceives a slightly different image. We also have precise information on the location of our eyes, the angle of our heads, and their distance to the ground surface. Our brains combine this information to create a three-dimensional scene. Our binocular ("two-eyed") vision means we can estimate the size, distance, and/or location of most objects - no further information needed.

However, almost all remote sensing systems have monocular ("one-eyed") vision,

which limits them to producing flat, two-dimensional imagery. Using the image alone, we cannot readily measure the size, distance, and location of objects in a scene, nor can we compare it with other images in that location - a must for earth observation applications! Much like the auxiliary information our brain uses to create a three-dimensional scene, we can make a two-dimensional image "spatially explicit" by measuring the following:

1. The precise location of the camera in three-dimensional space
2. The positioning or perspective of the camera

Finding the camera location is fairly straightforward. We can use a Global Positioning System (GPS) to record our exact coordinates. Depending on the platform - terrestrial, aerial, or spaceborne - we can use various tools to record the platform's height, altitude, and/or elevation.

The camera perspective, including lens angle and direction, heavily influences how objects are perceived in imagery. Similarly to how accidentally opening your phone camera on selfie mode is not ideal for a flattering photo of your face, there are favourable perspectives for observing different natural phenomena. It's therefore of high importance to carefully select the best perspective for the desired use of your imagery.

The precise angle of the camera is also crucial. Thinking back to map projections in Chapter 2, you will recall that representing our three-dimensional planet in a two-dimensional space causes certain regions to be heavily distorted in shape and size. You will also recall from earlier in this chapter how a camera's optical power changes the way objects at varying distances are seen.

There are four camera perspectives used for Earth observation discussed here: aerial, nadir (pronounced NAY-der), oblique and hemispherical. Each one is briefly explained below with photos and some example applications.

### 12.3 Aerial Perspective

The plane of the lens is perpendicular to the ground plane and the lens vector is pointed straight downwards at the ground. Figure 12.15 is an example.

Aerial imagery can be taken from remotely piloted aircraft systems (RPAs), airplanes, or satellites and thus has a huge range of resolutions and area coverage. It's highly sensitive to adverse weather, cloud cover or poor air quality, and variable lighting, so it needs to be carefully timed or collected at frequent intervals to account for unusable data.

Common applications: - *Mapping land cover and land use - Assessing ecosystem disturbance frequency and severity - Calculating indices such as Normalized Difference Vegetation Index (NDVI) and Normalized Difference Burn Ratio (NDBR) - Collecting climate and weather data - think thermal maps, storm tracking, coastline changes, etc*

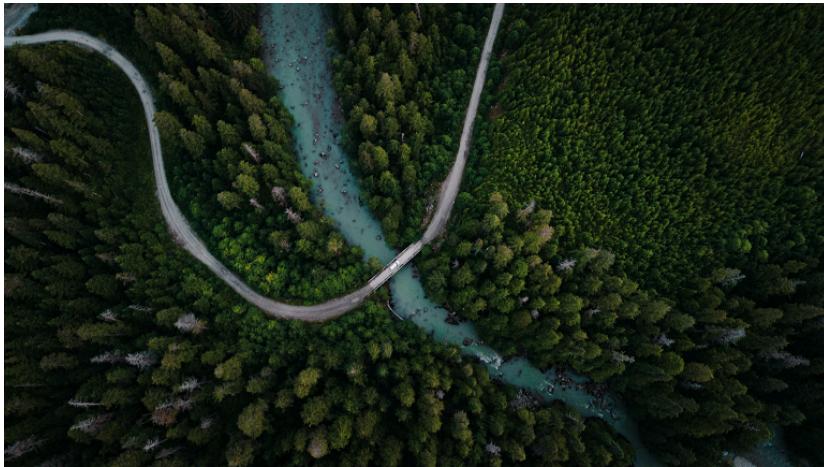


Figure 12.15: Aerial photo of forest, road, and river near Kitimat, BC [den Engelsen, 2020]. Unsplash License.

*It's important to note that the “ground plane” refers to a plane tangent to the geoid and not the physical ground surface. In variable terrain such as mountains, much of the ground will be seen “at an angle”, but the overall camera perspective is unchanged. See figures 12.16 and 12.17 for a visualization of what this looks like with regards to aerial imagery.*

## 12.4 Nadir and Zenith Perpsectives

When the focal plane of the lens is parallel to and pointed towards the vertical datum, then this perspective is known as **nadir**. The point opposite to nadir is the **zenith**, which is simply the location directly above nadir relative to the vertical datum (Figure 12.18). The imaginary line that connects the zenith and nadir points is usually perpendicular to the focal plane of a remote sensing system. In other words, the sensor is typically pointed straight up from the ground or straight down towards the ground. Any deviation from this is an oblique perspective, which is discussed in the next section.

Common applications: - Determining crown closure or canopy cover - Viewing branch networks - Measuring leaf area for the upper canopy - Astronomy and cosmological observation

## 12.5 Oblique Perspective

If the plane of the lens is *not* perpendicular to the vertical datum, then the imagery is considered to be **oblique**. Oblique imagery is ideally suited for comparing object sizes or viewing areas that would be otherwise occluded in

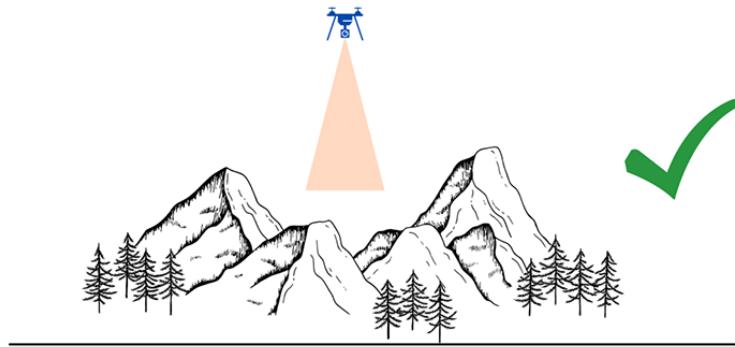


Figure 12.16: How aerial imagery should be taken. Claire Armour. CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/remote-sensing-systems.html#fig:12-aerial-good>.



Figure 12.17: How aerial imagery should NOT be taken. Claire Armour, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/remote-sensing-systems.html#fig:12-aerial-bad>.



Figure 12.18: Zenith perspective taken from the ground looking up to the canopy of an old growth tree on Vancouver Island, British Columbia. Pickell, CC-BY-SA-4.0.

aerial imagery. Nearly all terrestrial platforms take oblique imagery and it is readily used for airborne and spaceborne platforms. A scanning platform will have an oblique perspective when it is not at the nadir or zenith of its scan arc. Figure 12.19 is an example of an oblique image of a natural area.

Common applications: - *Viewing and measuring forest understorey and mid canopy* - *Assessing post-disturbance recovery* - *Assessing wildfire fuel loading* - *Providing context for aerial and nadir imagery* - *Comparing individual trees or vegetation*

## 12.6 Hemispherical Perspective

A hemispherical perspective has less to do with camera positioning and more to do with field of view, but it is still a “perspective”. It captures imagery in the half-sphere (hemisphere) directly in front of the lens. The radius of the hemisphere is dependent on the lens size and optical power of the hemispherical lens used in the camera. Figure X below visualizes a hemispherical perspective.

Due to the unusual shape of the lens, it captures a much larger proportion of a scene than we could normally take in without swiveling our heads or stitching photos into a mosaic, such as a panorama. A hemispherical lens will produce a circular rather than rectilinear output. The lens curvature will cause objects in the image to be highly distorted and, unlike rectilinear photos, cannot be



Figure 12.19: Oblique image of a forest harvest near Cold Lake in Saskatchewan.  
Ignacio San-Miguel, CC-BY-SA-4.0.

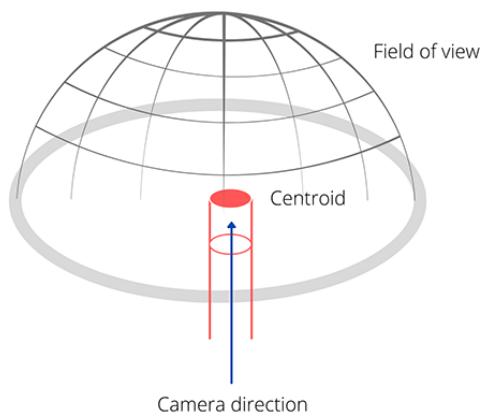


Figure 12.20: Visualization of a hemispherical perspective. Claire Armour, CC BY 4.0.

easily divided into pixels for analysis. However, hemispherical perspectives are uniquely suited to viewing large expanses of a scene all at once. For this reason, it is highly favourable for sports cameras, security cameras, and natural monitoring. Figures 12.21 and 12.22 shows hemispherical perspectives from two very different angles.



Figure 12.21: Hemispherical photo taken in the Bavarian forest [Wegmann, 2011], CC BY 3.0 Unported.



Figure 12.22: Picture of Newfoundland, Canada, taken by David Saint-Jacques during his space mission [Canadian Space Agency and NASA, 2019], CC BY 3.0 Unported.

Common applications - *Astronomy and cosmological observation - Tracking road and trail usage by wildlife, humans, and/or transport vehicles - Measuring Leaf Area Index (LAI) for the entire canopy*

## 12.7 Platforms

Remote sensing **platforms** are simply whatever a camera system or sensor is attached or affixed to. The platform can be stationary like a camera on a tripod or it can be mobile in the atmosphere or orbiting in space (Figure 12.23). The choice of platform can impact everything from the type of imagery that can be collected to the scale and frequency of the imagery. As a general rule of thumb, the farther you are above Earth, the more expensive the platform becomes, but the cost per area imaged is reduced drastically. For example, a handheld camera is relatively cheap compared with a multi-million dollar satellite, but it would cost a lot of time and resources to image large areas with a handheld camera compared with a satellite. Scale or resolution of your imagery will also tend to decrease the farther that you get from Earth. This means that each pixel in an image is representing a larger area on the Earth's surface. In the following sections, we will look at some examples and applications of various platforms for remote sensing systems.

### 12.7.1 Terrestrial Systems

Terrestrial platforms describe any platform that is near the ground surface. Usually terrestrial remote sensing systems are fixed and immobile where the sensors or camera systems are attached to a tower or a tree, but they may also be attached to vehicles such as the Google™ Street View vehicles that collect imagery from a 360 degree camera (Figure 12.24). One of the obvious limitations for mounting sensors to vehicles is that they are limited to traveling on roads, which limits what can be seen from the camera. But the clear advantage is that this is a cheap platform that only requires a driving licence to operate.

Many other terrestrial platforms are fixed or stationary, which means that they are either always observing the same feature or they might have a limited motorized range to pan from a fixed point. Phenological studies that aim to monitor the timing of different plant growth stages throughout the growing season such as budding, leaf-out and flowering will often use a stationary camera pointed towards the plant of interest. Figure 12.25 shows a time-lapse of images taken once per day at noon during the spring time near Grand Cache, Alberta. The changing leaf colour is clearly visible in the time lapse, which can be important for monitoring springtime wildfire risk [Pickell et al., 2017] or forage quality for wildlife such as grizzly bear [Bater et al., 2010].

Spectral responses from forest canopies can be monitored by radiometers such as the Automated Multiangular SPectro-radiometer for Estimation of Canopy reflectance (AMSPEC) instrument that can be mounted to a tower and sits high above the forest canopy (Figure 12.26). These “eyes above the forests” can provide important information about forest health and physiology. Since forest canopies are usually imaged from airborne and spaceborne platforms, these terrestrial observations provide a critical link for calibration with other imagery.

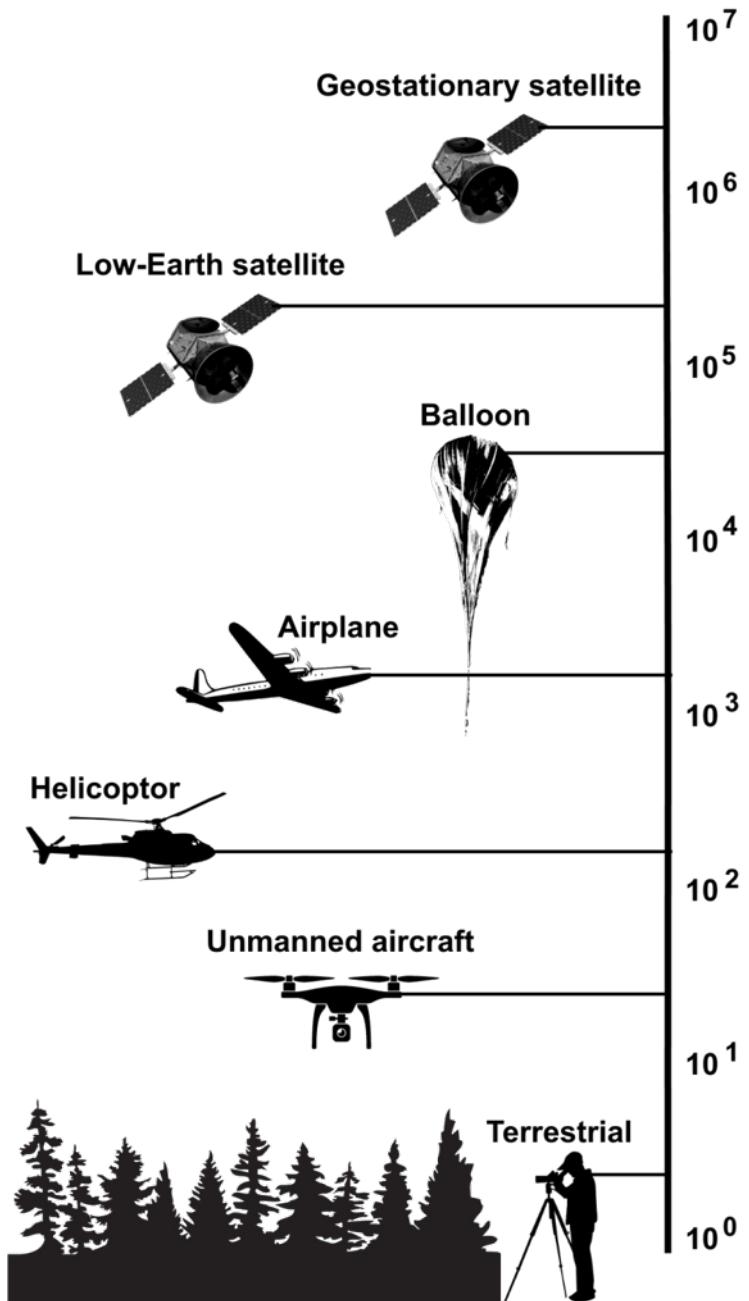


Figure 12.23: Different types of platforms for remote sensing systems. Image scale is represented on the y-axis in log scale. Pickell, CC-BY-SA-4.0.



Figure 12.24: A 360 degree camera mounted on a vehicle is used for collecting street view imagery for Google™ Maps [Leggett, 2014], CC-BY-SA-4.0.



Figure 12.25: Time lapse from a camera mounted on a tree. Each image is taken on a different day at noon [Bater]. CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/remote-sensing-systems.html#fig:12-phenological-camera-sequence>.



Figure 12.26: AMSPEC radiometer affixed to a carbon flux tower located in Buckley Bay, Vancouver Island, Canada [Coops]. Used with permission.

In Canada, there are significant stores of historical terrestrial imagery that were collected by the government for surveying the west. From the 1880's to as late as the 1950's, various government agencies of Canada collected over 5,000 terrestrial images of Canada's western territories and provinces primarily within the Rocky Mountains. Some of these locations have been re-imaged at the same terrestrial perspective during the modern era and show dramatic changes to the landscape such as glacial retreat and afforestation. Figure 12.27 shows the retreat of the Athabasca Glacier over nearly 100 years near the Wilcox Pass in Jasper National Park.

### 12.7.2 Aerial Systems

Historically, aerial platforms have played a major role in capturing remotely sensed imagery across Canada. Aerial systems were the first to achieve the bird's eye view and allow for large areas of sparsely populated Canada to be imaged in a standard way. The National Air Photo Library in Canada contains more than 6 million air photos across the country, some dating back to the 1920's. This historical archive consists of monochromatic, colour, and infrared imagery collected from fixed wing aircraft. Imagery can also be acquired from other types of aircraft including high altitude helium balloons, helicopters and even unmanned aerial vehicles (UAV).

One of the main reasons to acquire imagery from aircraft is the benefit of being able to image large areas at a relatively high spatial resolution. Most aerial photography can resolve objects between 1-10 cm on the ground. For forests, this means the ability to see branches and texture of the canopy, which can aid in identification of forest types and tree species. This resolution also allows for rare and relatively small ecosystems to be identified that would otherwise be obscured in satellite imagery. However, aerial systems are limited by the fact that they must be piloted under optimal weather conditions. Cost can also be prohibitive due to the need to pay for the aircraft and labour of a pilot.



Figure 12.27: Image pair of Athabasca Glacier from Wilcox Pass in Jasper National Park, Alberta, Canada. The historical image was taken in 1917 by A.O. Wheeler [Library and Archives Canada, 1917], and the modern image was taken nearly a hundred years later in 2011 by the Mountain Legacy Project [Mountain Legacy Project, 2011]. CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/remote-sensing-systems.html#fig:12-mountain-legacy-project>.

As a result, aerial images are usually not acquired very frequently or with any regularity.

Significant advancements have been made in recent years to reduce costs of aerial imagery through the use of UAVs. The benefits of a UAV system is that they are relatively cheap to operate, can be deployed rapidly in remote areas, and may be operated by a pilot with relaxed licensing and certification standards. However, UAVs are limited in the extent of the area that they may image due to battery life and the need for the aircraft to maintain a visible sight-line with the pilot on the ground. In Canada, UAVs are not permitted to operate within 5.6 km of airports and 1.9 km of heliports, which also limits their use in most urban areas. Some of the most advanced UAV systems can operate semi-autonomously and fly pre-planned routes and land before the battery drains down or if weather conditions are unsuitable.

Aircraft are subject to rotation along three axes (Figure 12.28). **Pitch** refers to rotation around the wings and controls whether the aircraft is ascending or descending. **Roll** refers to rotation around the fuselage (body) of the aircraft and controls which wing of the aircraft is higher than the opposite wing. **Yaw** refers to rotation around the vertical axis that is perpendicular to the fuselage and controls whether the aircraft is moving left or right. Each of these axes are important for understanding the conditions under which aerial photographs are acquired. The pitch and roll of the aircraft have perhaps the most pronounced effect on aerial imagery because any non-zero angle of pitch or roll (positive or negative) will produce an oblique image and cause scale to be inconsistent across the image. Because yaw is an axis that is perpendicular to the aircraft and also to Earth's surface, there is no impact from positive or negative yaw on image scale. However, large angles of yaw or roll can impact the ability to produce overlapping and adjacent stereo image pairs. Aerial photography is covered in more detail in Chapter 14.

### 12.7.3 Satellite Systems

Canada entered space in 1962 with the launch of the Alouette 1 satellite, the first country to launch a satellite after the Soviet Union and the United States. Since then, Canada has invested significantly into satellite-based remote sensing systems in order to monitor the vast and sparsely inhabited areas of the country. One of the primary advantages of remote sensing from satellite systems is the continuity of standard repeat images over large areas. These standards and imaging frequency are only made possible by a stable orbit and the autonomous nature of the satellite. All aerial systems are subject to atmospheric turbulence and therefore the quality of the imagery can depend on the rotation of the aircraft (pitch, roll, and yaw) and how well the human pilot can maintain the altitude, speed, and direction of the aircraft. By contrast, satellite systems can image continuously in a semi-autonomous mode, on a fixed orbit, at a relatively fixed speed and altitude. As a result, the sheer volume of images collected by satellite systems far exceeds any single aerial system. For example, the Landsat

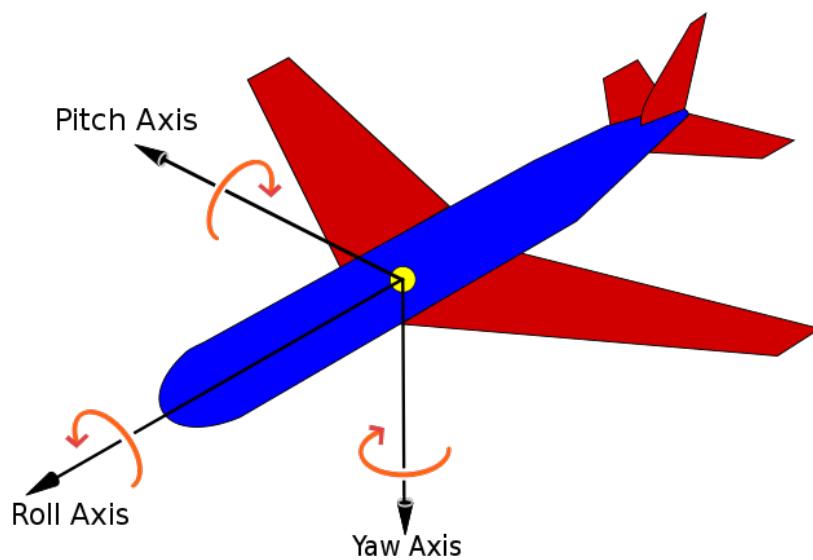


Figure 12.28: Pitch is the rotation of the aircraft over the axis of the wings, roll is the rotation of the aircraft over the axis of the fuselage, and yaw is the rotation of the aircraft over the axis of vertical axis that is perpendicular to the fuselage [Jrvz, 2010]. CC-BY-SA 3.0.

satellite program alone has collected over 10 million images from space since 1972 or approximately 555 images per day on average!

The primary limitation of satellite remote sensing systems is that imagery is not available prior to when Earth observing satellites were first launched in the late 1950's. Beyond the limitation of historical imagery, the main disadvantages to satellite systems are related to their orbits. Since satellites can not be easily maneuvered, it could take days before a particular satellite returns over some location of Earth to take an image. Some satellite systems have motorized sensors, which can be "tasked" in an off-nadir, oblique perspective. Other satellite systems are comprised of a constellation of copies of the same satellite and sensor to provide additional and more frequent coverage. The other limitations associated with satellites in orbit are the relatively low spatial resolution offered by space-based images (especially when compared with aerial systems that routinely achieve centimeter-level spatial resolution) and also the fact that space-based images are subject to atmospheric effects that can obscure the surface (e.g., clouds) or distort the reflectance of the ground surface (e.g., smoke, haze). The atmosphere is comprised of many aerosols and particles that can absorb, reflect or scatter electromagnetic radiation at different wavelengths. Thus, satellite systems are limited to observing only the wavelengths that can be transmitted through the atmosphere, known as **atmospheric windows**.

In the following sections, we will look at orbits and their role in different satellite systems and then turn to look at some important satellite systems for environmental management.

## 12.8 Orbital Physics

**Orbits** are curved paths around a celestial object like the Earth or the Moon. Sir Issac Newton observed three important Laws of Motion that are relevant for describing how orbits work: - The First Law of Inertia: an object will remain at rest or a constant speed of motion until acted upon by another force - The Second Law of Acceleration: the acceleration of an object depends on its mass and the amount of force applied - The Third Law of Action and Reaction: every force acting on an object has an equal force that reacts in the opposite direction

The First Law says that anything moving through space will continue to move through space at a constant speed forever (even if that speed is 0) unless another object exerts some force on it. The Second Law says that objects can accelerate with force, but more massive objects require more force. Finally, the Third Law constrains the other two with the fact that every interaction between two objects causes two forces to occur in opposite directions.

In the simplest terms, orbits form when two objects are in motion near each other in the vacuum of space. For satellites to reach orbit around Earth, they must accelerate at very high speeds to escape Earth's gravity and maintain their

inertial motion (Second Law). At specific speeds, satellites can maintain their motion and continue to interact with Earth, exerting a small force on Earth that is reciprocated (Third Law). The exact path of an orbit is a function of the mass and gravitational acceleration between the two objects. In other words, objects orbit each other and are in constant free fall towards the other due to the Third Law. For example, the Earth is falling towards the Moon and the Moon is falling towards the Earth. Both objects are orbiting around the same imaginary point representing the center of mass of both objects. The same physics apply to the orbits of satellites around Earth. Since Earth has much more mass than any artificial satellite, the center of mass between Earth and any satellite is generally near the center of Earth itself, thus we see the illusion that the satellite “orbits” Earth.

Consider Newton’s cannonball thought experiment: you fire a cannonball perpendicular to Earth’s surface at the top of a tall mountain. If the cannonball is traveling at  $0 \text{ km} \cdot \text{s}^{-2}$ , then the cannonball falls to the ground due to the force of gravity. If the cannonball is traveling at  $1 \text{ km} \cdot \text{s}^{-2}$ , then the cannonball travels some distance over an arc before it eventually falls back to Earth due to gravity (Figure 12.29 A). If you fire the cannonball at a slightly faster speed of  $3 \text{ km} \cdot \text{s}^{-2}$ , then the cannonball travels a distance farther than before but still eventually falls back to Earth due to gravity (Figure 12.29 B). If the cannonball is traveling at least  $7.8 \text{ km} \cdot \text{s}^{-2}$ , then the speed of the cannonball is roughly equivalent to the force of gravity that is trying to pull the cannonball back to Earth and therefore the cannonball maintains an approximately circular orbit around Earth (Figure 12.29 C). In other words, the velocity of the cannonball is faster over Earth than it is falling towards Earth, which results in a curved path or orbit. Fire the cannonball at any faster speed and you can achieve elliptical orbits (Figure 12.29 D) or even orbits that escape Earth’s gravity altogether (Figure 12.29 E).

Another important force is drag, which is the equal force applied in the opposite direction to acceleration (Third Law). Drag is important for satellites to reach orbit and stay there. Earth’s atmosphere exerts an equal and opposite force to the direction the space vehicle is leaving the launchpad and also creates friction that can slow down satellites that are near the transition between the atmosphere and space. Thus, in order for satellites to maintain orbit, they must travel at high speeds and also high altitudes above the ground to avoid the force of drag. Generally, the speed of a satellite is inversely related to the altitude. So satellites that are closer to Earth must travel at much faster speeds than satellites that are farther from Earth. In the following sections, we will look at some examples of different types of orbits and their role in specific remote sensing systems.

### 12.8.1 Low Earth Orbit (LEO)

**Low Earth Orbit (LEO)** is a critical entry point to space because it marks the transition between Earth’s upper atmosphere and the vacuum of space. Many

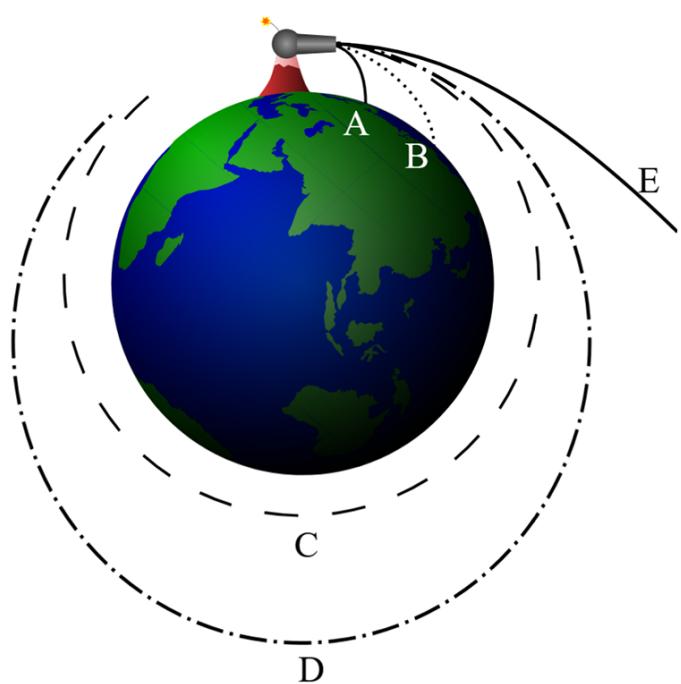


Figure 12.29: Netwon's Cannonball thought experiment. Brian Brondel, CC-BY-SA-3.0.

Earth-observing satellites are placed in LEO between 200-2,000 km altitude above Earth and this region comprises Earth's thermosphere and exosphere layers of the atmosphere. At the extreme of this range, Earth's atmosphere becomes so rarefied that individual atoms of hydrogen and helium can travel hundreds of kilometers without encountering another atom. Many of these atoms will be swept away by solar winds into the depths of space and the density of the atmosphere is so low that it is treated as a vacuum. Thus, drag on a satellite from Earth's atmosphere is practically nonexistent at these altitudes.

Relatively speaking, LEO is the most crowded region of near-Earth space due to decades of space vehicle launches and rare satellite collisions that have left behind debris, small particles, and whole components of past space vehicles. As a consequence, it is also the most dangerous region of near-Earth space because even the smallest space debris can be traveling at orbital speeds of up to  $14 \text{ km} \cdot \text{s}^{-2}$ . Significant care is taken to track and model space debris because all satellite and human space craft must navigate LEO in order to reach higher orbital altitudes.

### 12.8.2 Near-Polar and Sun-synchronous Orbits

**Near-polar or sun-synchronous orbits** are a special type of LEO where the satellite follows a path that travels approximately from pole to pole. The special property for this orbit is that the orbital period of 96-100 minutes (the time needed to complete a full orbit around Earth) is approximately equivalent to the timing of Earth's rotation. This results in the satellite crossing the Equator (or any other parallel on Earth, depending on the inclination) twice at the same local time, once during the day and once during the night. This synchronization with the Sun is very important for many passive Earth observing satellites that require consistent illumination conditions from image to image.

### Recall This

Earth is an oblate ellipsoid with a slightly shorter axis of rotation than the equatorial axis, so a polar orbiting satellite with a near perfect circular orbit can vary in "altitude" by as much as 30 km.

### 12.8.3 Medium Earth Orbit (MEO)

**Medium Earth Orbit (MEO)** occurs at altitudes between 2,000-35,786 km or orbital periods more than 2 hours and less than 24 hours. This region of space is much less crowded compared with LEO and satellite activity is primarily characterized by navigation and communication services. Satellites in MEO are traveling at nearly half the speed ( $\sim 4 \text{ km} \cdot \text{s}^{-2}$ ) compared with LEO satellites ( $7.8 \text{ km} \cdot \text{s}^{-2}$ ) and can therefore remain above the visible horizon of Earth for several hours, which is what makes satellite communication, TV broadcasts, and navigation possible. For example, Global Navigation Satellite Systems (GNSS)

such as the Global Positioning Service (GPS) use a constellation of satellites in MEO that are oriented with different inclinations relative to Earth's Equator to ensure that several satellites are always in view for nearly any location on Earth (see Chapter 4 for more on GNSS).

#### 12.8.4 Geosynchronous Equatorial Orbit (GEO)

Were you wondering why MEO extends to such an exact number at the extreme altitude of 35,786 km? That is because 35,786 km is the distance from Earth at which the orbital period of a satellite at the Equator is equivalent to Earth's rotational period of 24 hours. This is known as **Geosynchronous Equatorial Orbit (GEO)** because this orbit only occurs directly above the Earth's Equator. Satellites in GEO are **geosynchronous**, meaning they are always visible in the same location of the sky no matter the time of day or the season. For this reason, these orbits are sometimes referred to as **geostationary**. The advantage of geostationary orbit is that the satellite can continuously image the same visible portion of Earth 24 hours a day. Thus, nearly all weather and communication satellites are in geostationary orbit, allowing transmissions to be relayed across a network of geostationary satellites and ground antenna like a ping-pong ball.

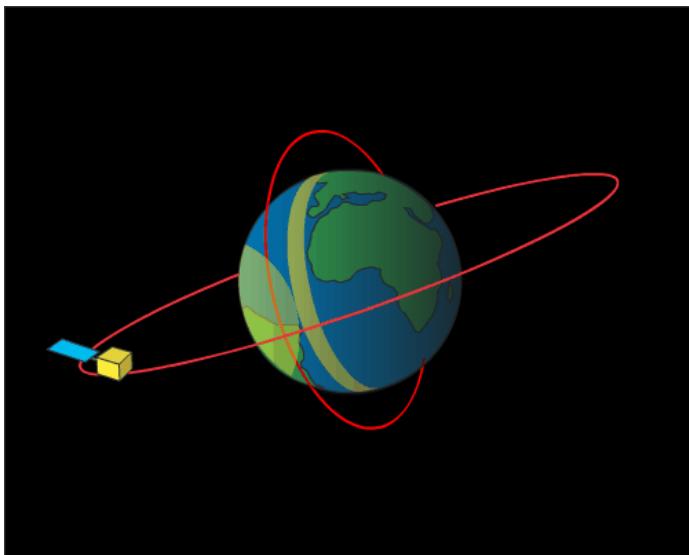


Figure 12.30: Comparing Sun-synchronous and geosynchronous orbits. The yellow area shows the portion of Earth's surface that is visible during a single orbit. Notice that Sun-synchronous orbit observes at a consistent local time while the geosynchronous orbit observes a constant location. Credit: NOAA/JPL-Caltech. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/remote-sensing-systems.html#fig:12-sunsynchronous-geosynchronous-orbits>.

## 12.9 Summary

### Reflection Questions

1. What other “remote sensing systems” can you think of that you use day-to-day?
2. How do you define space? How far is space from the ground?
3. Search the web for any satellite or remote sensing system discussed in this chapter. What applications or research did you find?

### Practice Questions

1. How many degrees of motion does the Enhanced Thematic Mapper + (ETM+) aboard the Landsat 7 satellite have?
2. The visible band of the Advanced Very High Resolution Radiometer (AVHRR) has an instantaneous field of view (IFOV) of  $1.300 \times 10^{-3}$  radians (rad). The satellite that carries AVHRR currently orbits Earth at a nominal altitude of 833 km. What is the spatial resolution at nadir for this band?
3. AVHRR is an example of a whisk broom scanning system that uses a scanning mirror to reflect radiation into a single detector, one pixel at a time across the track of the orbit. If the system has a scan angle of  $\pm 55.37$  degrees from nadir, what is the approximate swath width of the visible band?

The Operational Land Imager (OLI) on board Landsat 8 is an example of a push broom system. There are a total of 6,916 detectors on the OLI, but the sensor is designed so that the detectors are staggered in a butcher block pattern across the focal plane, which ensures a  $15^\circ$  field of view without any moving parts. Due to this design, the blocks of linear detectors overlap slightly to create a gap-free swath width of 185 km with a spatial resolution of 30 m.

4. What is the approximate instantaneous field of view for the OLI detectors in degrees?
5. What is the necessary flying altitude of the spacecraft to maintain the swath width and spatial resolution?

## **Chapter 13**

# **Image Processing**

Written by Paul Hacker and Paul Pickell

The collection of imagery is challenging and time consuming, with sensor design and deployment often requiring the development of new technologies. Driven by the growing demand for relevant information about environmental change, great emphasis is placed on the creation of “the next new satellite” or “a smaller, yet more powerful drone”. Although innovation in these physical technologies is critical for the advancement of image collection, they don’t guarantee that the imagery will be useful. In reality, the pixel values collected by a sensor are not purely the reflectance values from the surface of interest. Depending on the wavelengths being observed, there may be variations in the amount of photons emitted from the light source and a variety of atmospheric effects, such as scattering. There may also be slight inconsistencies in images collected by the same sensor on the same day or adjacent areas, which would affect the quality of analyses.

### **Learning Objectives**

1. Relate common issues of image collection to relevant image processing techniques
2. Understand the logic supporting the use of specific image processing techniques
3. Explore a variety of processing methodologies employed by published research.

### **Key Terms**

pixel, spatial resolution, temporal resolution, radiometric resolution, spectral resolution, geometric correction, atmospheric correction,

## 13.1 Overview

The application of correction methods that address these inconsistencies are often described generally as “image processing”. In this chapter we will explore a variety of common image processing techniques and strive to understand the logic behind employing one, or more, to remote sensed imagery. Before diving into specific processing workflows that render imagery scientifically useful, however, it is important to review some key terms.

First and foremost, **image processing**, or **digital image analysis**, refers to any actions taken to improve the accuracy of one or more component of raw imagery. In remote sensing science, the goal of image processing is to generate a **product** that provides accurate and useful information for scientific pursuit. This is in contrast to image processing for artistic purposes, which could include many similar steps, but focus on generating a product that is visually appealing.

**Noise** is another common term associated with image processing and it refers to any element of the data that is not wanted. There are a variety of noise types, which we will discuss later. In contrast to noise, **signal** describes wanted components of the imagery. Combined, signal and noise provide guidance on what specific steps should be taken in relation to the data during image processing. Before they are addressed, however, it is also important to confirm the spatial accuracy of the data.

It is also important to review the elements of an **image**, or **raster**. The term raster refers to a data type comprised of a number of cells to which values are assigned. An image is a raster with cells values that represented some function of observed electromagnetic radiation. The raster cells in images are often referred to as \*picture elements, **or** pixels\*\*. The terms cell and pixel are considered interchangeable when discussing imagery. An empty raster, or image, would contain pixels with no information. To say an image has been collected would simply mean that a sensor has collected information and stored that information in adjacent pixels. Although this seems straight forward there are a multitude of environmental and engineering factors that can affect the accurate collection of information. It is these confounding factors that image processing attempts to resolve in hopes of generating a data product that can be compared across space and time.

## 13.2 Geometric Correction

By definition, remotely sensed data is collected by a sensor at some distance apart from the object(s) being observed. The design of a sensor generally determines the distance at which the desired observation is to be made. For example, the camera in a cell phone is designed for relatively close range observations of electromagnetic radiation, while a sensor mounted on a satellite to operate relatively far from the target objects.

In many instances, the sensor is in motion and in many cases is subject to influence from environmental factors like wind. Interactions with phenomena like wind can cause the instantaneous field of view (IFOV) to move slightly during data collection, introducing spatial errors to the imagery. On top of issues relating to sensor displacement, sensors in motion may also observe adjacent areas with different topographic properties.

A combination of these two effects could be visualized by imagining an airplane flying over a forested hillside. As the sensor collects data the plane can be buffeted with wind, changing the direction of a sensor's IFOV to a location that is not the original target. On top of issues with sensor movement, the elevation of the ground is constantly undulating, altering the distance at which the sensors observes the landscape below. These two issues that affect the spatial components of image collection compromise the accuracy of an image and need to be corrected. The general term used to refer to the spatial correction of an image is geometric correction.

### 13.3 Orthoimagery

An orthoimage is an aerial photograph or satellite imagery geometrically corrected so that the scale is uniform. Unlike orthoimages, the scale of ordinary aerial images varies across the image, due to the changing elevation of the terrain surface (among other things). The process of creating an orthoimage from an ordinary aerial image is called orthorectification. Photogrammetrists are the professionals who specialize in creating orthorectified aerial imagery, and in compiling geometrically-accurate vector data from aerial images.

Digital aerial photographs can be rectified using specialized photogrammetric software that shifts image pixels toward or away from the principal point of each photo in proportion to two variables: the elevation of the point of the Earth's surface at the location that corresponds to each pixel, and each pixel's distance from the principal point of the photo. Aerial images need to be transformed from perspective views into plan views before they can be used to trace the features that appear on topographic maps, or to digitize vector features in digital data sets.

Compare photographs in Figure ref(fig:13-ortho). Both show the same gas pipeline, which passes through hilly terrain. Note the deformation of the pipeline route in the photo on the left relative to the shape of the route on the orthoimage to the right. The deformation in the photo is caused by relief displacement. The original photo would not serve well on its own as a source for topographic mapping.

Think of it this way: where the terrain elevation is high, the ground is closer to the aerial camera, and the photo scale is a little larger than where the terrain elevation is lower. Although the altitude of the camera is constant, the effect of the undulating terrain is to zoom in and out. The effect of continuously-varying

scale is to distort the geometry of the aerial photo. This effect is called relief displacement.

### 13.4 Relief Displacement

An important component of geometric correction deals with the effects of elevation on the pixels in an image. Changes in the terrain over which an image is collected lead to inconsistencies in the distances at which information is collected. These differences lead to objects in the image appearing in location inaccurate with reality and can be rectified using the spatial information of the sensor, datum and object. Examples of this effect can be observed in any photograph containing tall structures, which would appear to be leaning outward from the center of the image, or principal point. The rectification of relief displacement can be represented by the equation:

$$d = rh/H \quad (13.1)$$

where  $d$  = relief displacement,  $r$  = distance from the principal point to the image point of interest,  $h$  = difference in height between the datum and the point of interest and  $H$  = the height of the sensor above the datum.

### Your turn!

What is the image displacement of a pixel that is 0.5 mm from the principal point, 57 m below the datum, and collected from a sensor that is 135 m above the datum?

### 13.5 Georeferencing

The removal of inaccuracies in the spatial location of an image can be conducted using a technique called georeferencing. The basic concept of georeferencing is to alter the coordinates of an image through the association of highly precise coordinates collected on site using a GPS. Coordinates collected in the field are often called ground, or control points, and form the base of successful georeferencing. In general, increasing the amount and accuracy of ground points leads to increased spatial accuracy of the image.

There are a variety of techniques used to transform the coordinates of an image based on control points, all of which require some level of mathematics. The complexity of the polynomials used to transform the dataset, the more accurate the output spatial coordinates will be. Of course, the overall accuracy of the transformation depends on how accurate the control points are. Upon transforming a rasters coordinates it is important to evaluate how accurate the output raster is compared to the input raster. A common method of evaluating

the success is through calculating the square root of the mean of the square of all error, often called the Root Mean Squared Error (RMSE). In short, RMSE represents the average distance that the output raster is from the ground or control points. The smaller the RMSE, the more accurate the transformation.

## 13.6 Georegistration (georectification)

Similar to georeferencing, georegistration involves adjusting the raw coordinates of an image to match more accurate ones. In the case of georegistration, however, ground points collected in the field are replaced with coordinates from a map or image that has been verified as spatially accurate. This method could be considered a matching of two products, enabling the two products to be analyzed together. An example would be matching two images collected one year apart. If the first image is georeferenced accurately the second image can simply be georegistered by identifying shared features, such as intersections or buildings, and linking them through the creation of control points in each image.

## 13.7 Resampling

Despite all the efforts to accurately place an image in space, it is likely that any spatial alterations also change the shape or alignment of the image's pixels. This dislocation between pixel sizes within the image, as well as potential changes in their directionality alter the capacity to evaluate the radiation values stored within them. Imagine that a raw image is represented by a table cloth. On top of the table cloth are thousands of pixels, all of uniform height and width. The corners of this table cloth each have X and Y coordinates. Now imagine that you have to match these coordinates (the corners) to the four more accurate ground points that don't match the table cloth. Completion of this task requires you to stretch one corner of the table cloth outwards, while the other three corners are shifted inside, leading the table cloth to alter its shape and, in doing so, altering the location and shape of the raw pixels.

This lack of uniformity in pixel size and shape could compromise future image analysis and need to be corrected for. To do so, users can implement a variety of methods to reassign the spatially accurate values to spatially uniform pixels. This process of transforming an image from one set of coordinates to another is called resampling [Parker et al., 1983].

The most important concept to understand about resampling is also the first step in the process, which is the creation of a new, empty raster in which all cells are of equal size and are aligned North (Insert demo image). The transformed raster is overlaid with the empty raster with a user defined cell size before each empty cell is assigned a value based on values from the transformed raster. Confusing? Perhaps, but there are a variety of processes that can be used to assign cell values to the empty raster and we will discuss three in hopes of clarifying

this methodology.

### 13.8 Nearest Neighbor

Nearest neighbor (NN) is the most simple method of resampling as it looks only at one pixel from the transformed raster. This pixel is selected based on the proximity of its center to the center of the empty cell and the value is added without further transformation. The simplicity of this method makes it excellent at preserving categorical data, like land cover or aspect, but struggles to capture transitions between cells and can result in output rasters that appear somewhat crude and blockish.

### 13.9 Bilinear Interpolation

In contrast to NN resampling, bilinear interpolation (BI) uses the values of multiple neighboring cells in the transformed raster to determine the value for a single cell in the empty raster. Essentially, the four cells with centers nearest to the center of the empty cell are selected as input values. A weighted average of these four values is calculated based on their distance from the empty cell and this averaged value becomes the values of the empty cell. The process of calculating an average means that the output value is likely not the same as any of the input values, but remains within their range. These features make BI ideal for datasets with continuous variables, like elevation, rather than categorical ones.

### 13.10 Cubic Convolution

Similar to bilinear interpolation, cubic convolution (CC) uses multiple cells in the transformed raster to generate the output for a single cell in the empty raster. Instead of using four neighbors, however, this method uses 16. The idea supporting the use of the 16 nearest neighbors is that it results in an output raster with cell values that are more similar to each other than the values of the input raster. This effect is called smoothing and is effective at removing noise, which makes CC the ideal sampling method for imagery. There is one drawback of this smoothing effect, however, as the output value of a cell may be outside the range of the 16 input cell values.

### 13.11 Atmospheric Correction

Following similar logic to that promoting the need for geometric correction, atmospheric correction is intended to minimize discrepancies in pixel values within and across images that occur due to interactions between observed radiation and atmosphere. The severity of impact that the atmosphere has on the observation

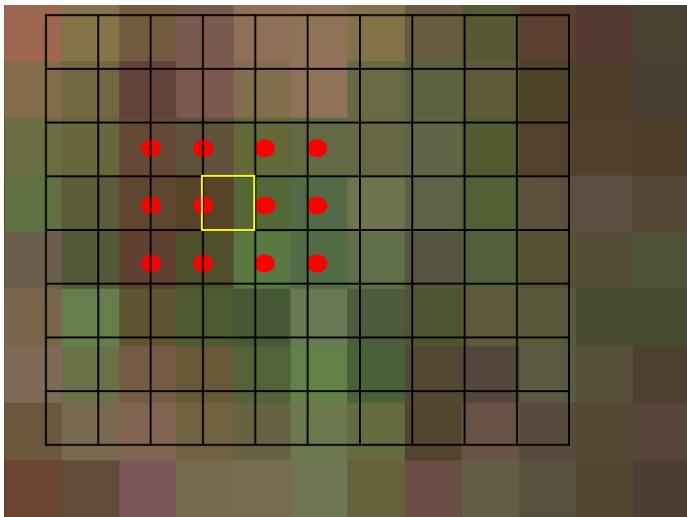


Figure 13.1: Demonstration of neighbour selection (red) using cubic convolution resampling to determine the value of a single cell (yellow) in an empty raster. Hacker, CC-BY-4.0.

of electromagnetic radiation relates to changes in the atmosphere during collection and the specific wavelengths being collected. The majority of impacts are caused by the three main types of scattering, which were presented in Chapter [ref\(fundamentals-of-remote-sensing\)](#).

## 13.12 Atmospheric Windows

A key characteristic of the earth's atmosphere that impacts the collection of passive remotely sensed data is the impediment of certain wavelengths. If solar radiation of a specific wavelength cannot reach Earth's surface, it is impossible for a sensor to detect the radiance of that wavelength. There are, however, certain regions of the electromagnetic spectrum (EMS), called atmospheric windows, that are less effected by absorption and scattering than others and it is the observation of these regions that remote sensing relies on. Upon reaching the Earth's surface, however, there are a variety of atmospheric constituents that can affect image quality.

## 13.13 Clouds and Shadows

Two of the most common culprits in the disturbance of remotely sensed imagery are clouds and shadows. Both are relatively transient, making the prediction of their inclusion in an image difficult and rendering their effects within an image relatively inconsistent. On top of issues of presence, each introduces unique

challenges for image correction.

Clouds are an inherent component of Earth's atmosphere and therefore should warrant respect and care in image processing, rather than sighs of frustration. You could imagine that a single image with 30% cloud cover may not be entirely useful, but their aforementioned permanent transience means that data users must work to reduce their effects, if not remove them entirely.

When approaching the removal of clouds, it is important to recall the physics that drive Mie scattering (Chapter ref(fundamentals-of-remote-sensing)). Essentially, water vapors in the atmosphere scatter visible and near infrared light and generate what appears to be white objects in the sky. Since the visible and near infrared regions of the EMS fall within an atmospheric window in which many sensor detect radiation, clouds can be recorded as part of an image.

The removal of clouds is often referred to as masking and can prove challenging depending on the region in question as they also generate cloud-shadows. You could imagine a study attempting to evaluate snow cover over a landscape using imagery comprised of wavelengths in the visible region of the EMS. If there was intermittent cloud cover, cloudy areas with no snow could be classified as having snow and snowy areas with cloud-shadows may be classified as no snow.

Essentially, the removal of clouds utilizes the fact that clouds are cooler than the Earth's surface and can be identified using thermal data. Albedo, which is the capacity of a surface to reflect measured from 0 (no reflectance) to 1 (full, pure white reflectance), can also be calculated and used to identify the significantly brighter clouds. Once the clouds are identified and the mask is created, it is possible to identify and remove cloud-created shadows.

Shadows present unique problems for image analysis as they can shade out underlying structures and also be classified as separate, individual objects. The former issues presents problems for studies evaluating land cover, while the latter confounds machine learning algorithms attempting to identify unique classes in the image based on spectral similarities. Another confounding issue is that the location and size of shadows change throughout the day in accordance with the sun.

An important feature of any shadow is that the area shaded is still considered to be illuminated, but only by skylight . The exclusion of sunlight from the area creates a unique opportunity for shadow identification and removal [Finlayson and Hordley, 2001]. Finlayson and Hordley's method of shadow removal is complex, using derivative calculus to capitalize on the fact that a illumination invariant function can be recognized based solely on surface reflectance. Although more complicated than Martinuzzi et al.'s approach, it may be worth reviewing Finlayson's work if you are interested in learning more about shadow removal.

## 13.14 Smoke and Haze

Smoke and haze present unique issues to image processing as they tend to vary in presence, consistency and density. They also represent different types of scattering, with smoke causing Mie scattering and haze causing non-selective scattering. Makarau et al. demonstrated that haze can be somewhat removed through the creation of a haze thickness map [Makarau et al., 2014]. This methodology is equally as complex as that of Finlayson's and is perhaps beyond the scope of this book. It is important to note, though, that removal of shadows, clouds, smoke and haze relies on an understanding of how their respective scattering types affect incoming solar radiation. Successful removal, then, depends on understanding which spectral bands relatively unaffected by the particular type of scattering occurring within the image.

## 13.15 Radiometric Correction

We have discussed how the creation of an image by a remote sensor leads to slight variations in spatial and atmospheric properties between pixels and that these inconsistencies must be corrected for. In this section, we will discuss some issues affecting the information within a pixel and some common remedies. In essence, we will explore how the raw digital numbers collected by a sensor can be converted to radiance and reflectance.

## 13.16 Signal-to-noise

A key concept of radiometric correction is the ratio of desired information, or signal to background information (noise) within a pixel. The signal-to-noise ratio (SNR) is a common method of presenting this information and provides an overall statement about image quality. A common method of calculating SNR is to divide the mean ( $\mu$ ) signal value of the sensor by its standard deviation ( ), where signal represents an optical intensity. (Equation (13.2))

$$SNR = \text{signal}/\sigma_{\text{signal}} \quad (13.2)$$

It is clear, through Equation (13.2), that the average signal value of an instrument represents the value that its designers desire to capture. It is also clear that an increase in signal leads to an improved SNR. What remains unclear, however, is what causes a sensor to observe and record undesired noise to be recorded. In reality, there are a variety of noise types that can affect the SNR of a sensor.

### 13.17 Readout Noise

Readout noise is created through the inconsistencies relating to the interaction of multiple physical measurement electronic devices. Since it is impossible to have a sensor without physical devices, readout noise is inherent in all sensors. Readout noise is therefore equal to any difference in pixel value when all sensors are exposed to identical levels of illumination. There are a variety of technical methods used to correct for this error, but the concepts and mathematics supporting them are perhaps beyond the scope of this book. Check out *Readout Noise* by Michael Richmond if you are interested in learning more about readout noise.

### 13.18 Thermal Noise

Another inherent type of sensor noise is thermal noise. Thermal noise occurs in any device using electricity and is caused by the vibrations of the devices charge carriers. This means that thermal noise can never fully be removed from an image, although it can be reduced by lowering the temperature of the environment at which the sensor is operating.

### Your turn!

Calculate SNR or it's associated values for various Landsat sensors:

1. OLI = signal 5288.1, standard deviation - 18.7
2. TM = : 0.4,  $\mu$ : 5.8
3. ETM+ = SNR: 22.3,  $\mu$ : 13.4

### 13.19 Case Study: Title of Case Study Here

You see textual case study content here

### An overview of Landsat Processing

The field of remote sensing has witnessed the creation of a variety of national programs designed to observe the Earth's surface. Landsat is one of these programs and has enabled the collection of terrestrial information from space for over 40 years. Initiated by NASA, Landsat is now run by the United States Geological Survey (USGS) who provide the data they collect to users for free. The imagery collected from Landsat sensors has proven to be highly useful for environmental monitoring and the mission is scheduled to continue into the future.

Despite its success, however, the imagery collected by Landsat sensors continues to face the same processing issues as most other sensors. Variations in geometric and atmospheric characteristics exist within and across images, and radiometric inconsistencies also occur. To combat these issues, the USGS has implemented a tiered image processing structure.

Level-1 processing is designed to address the geometric and radiometric inconsistencies present in an image. The USGS utilizes a combination of ground control points (GCP), digital elevation models (DEM) and internal sensor calibrations to correct an image. Within Level 1 there are a variety of sub-levels that users can select from, each varying based on the amount of correction that has been completed. USGS's informational page [Landsat Level-1 Processing Details](#) is a good resource for learning more about the specifics of Landsat Level-1 processing.

With the geometric and radiometric correction complete in Level-1, Level-2 processing is designed to reduce atmospheric inconsistencies. The output images created from Level-2 processing are considered to be Science Products (images with corrections completed at a quality suitable for scientific inquiry without further processing) and accurately represent surface reflectance and surface temperature. Data regarding surface reflectance is useful for evaluating a variety of environmental metrics, such as land cover, while surface temperature lends insight to vegetation health and the global energy balance. It is important to note that there are two Collection levels at which Landsat processing takes place, with Collection 2 representing the highest quality processing stream. You can explore the differences between collections of Landsat Collection 1 vs 2 imagery [PDF], while information about Collection 2 images processed at a Level-2 standard can be observed on USGS's [Landsat Collection 2 Level-2 Science Products](#) information page.

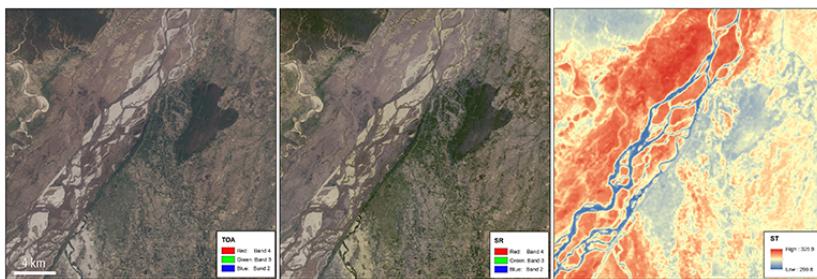


Figure 13.2: Three unique Landsat 8 Collection 2 images. In order from left to right: Level-1 Top of Atmosphere reflectance image (no atmospheric correction), Level-2 atmospherically corrected surface reflectance and Level-2 surface temperature. Images were collected on May 3, 2013 over the Sapta Kosh River in Bairawa, Nepal [Bouchard and US Geological Survey, 2013]. Public domain.

As you may realize, there are a variety of processing options that Landsat users

can select from. Each of these Collections and Levels present different opportunities for scientific study and allow users to customize processing streams based on the needs of their project. Such flexibility is key for the continued improvement of Landsat products and promotes the use of Landsat imagery across a broad range of users. It is important to remember, however, that the selection of pre-processed data for scientific inquiry requires the user to understand the foundations upon which correction were made. Be sure to draw on the fundamentals learned in this chapter to evaluate the usefulness of any processed data you consider in your work.

## 13.20 Image Enhancement

So far, in this Chapter, we have discussed methods of correcting spatial, atmospheric and radiometric errors that are commonly present in remotely sensed images. While the removal of these artifacts is necessary, it is also important to explore some common methods of enhancing an image once these aforementioned corrections have been made. Both image stretching and sharpening have roots in spatial and radiometric correction, so keep your mind open to the inherent links that arise.

### 13.21 Stretching

Image stretching refers to the adjustment of radiometric values of the input methods to better exploit the radiometric resolution of an image. In principle, the distribution of radiometric values within a image is altered in a manner that improves its capacity to perform a desired task (). For instance, if an image collected with an 8-bit radiometric resolution (256 radiometric values; 0-255). appears too dark, it is likely that the distribution of pixel values is centered on a radiometric value greater than 127 (middle of a 8-bit scale). In fact, of 0 represents white and 255 represents black, it is very likely that the majority of pixels are closer to 255. In this case, the lowest value(s) observed in the image can be adjusted to 0, the relationship of this change can be determined and then applied to all other observed values.

### 13.22 Smoothing

Image smoothing is a process used to reduce the noise in an image. Essentially, a filter of specific window size is passed over surrounding cells and an output value for the center cell is determined by a pre-defined algorithm. Similar to the windows used for resampling, smoothing window sizes can be user defined to generate images that are useful for specific applications. Despite its usefulness as a noise reduction technique, smoothing can also have negative effects on image quality through the loss of detail.

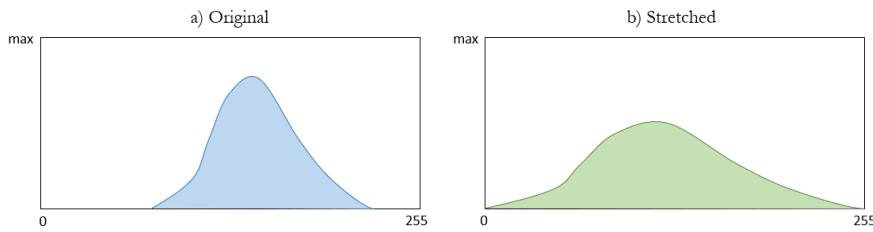


Figure 13.3: Example of how (a) the original distribution of radiometric values in a image is (b) stretched. Hacker, CC-BY-4.0.

## Call out

If you are interested in learning more about how sharpening can impact hyperspectral imagery you can check out [Inamdar et al., 2020]. Their research demonstrates that recorded pixel values contain information from areas beyond the traditional spatial boundary of a cell. These findings have interesting implications for a variety of applications.

## 13.23 Summary

This chapter provided an overview of common image processing techniques and discussed the logic that supports their usage. Overall, each technique strives to create imagery that is consistent across time and space in order for individual pixel values to be evaluated and/or compared. Although necessary, these processes can take time and need to be applied in accordance with the desired application. Understanding the general workflow of image processing will allow you to determine what steps should be taken to create the highest quality imagery for your research.

## Reflection Questions

1. Define geometric correction and discuss one of its components.
2. List three resampling techniques and describe the differences.
3. What is the differences between cloud and smoke with regards to scattering?
4. Why is the signal-to-noise ratio important for evaluating image quality?



## **Chapter 14**

# **Image Analysis**

Written by Paul Pickell

Introduction here.

## **Learning Objectives**

1.

## **Key Terms**

Term

## 14.1 Aerial Photography and Photogrammetry

14.1.1 Pitch, Roll, and Yaw

14.1.2 Stereo Vision

14.1.3 Shape

14.1.4 Pattern

14.1.5 Size

14.1.6 Tone and Colour

14.1.7 Shadow

14.1.8 Texture

14.1.9 Association and Context

## 14.2 Image Classification

14.2.1 Land Cover Versus Land Use

14.2.2 Supervised Classification

14.2.3 Unsupervised Classification

14.2.4 Classification Algorithms

## 14.3 Time Series Analysis

14.3.1 Change detection

14.3.2 Detecting spectral trends

14.3.3 Inferring surface activity

## 14.4 Case Study: Sea Ice Change Analysis in the Beaufort Sea

*Case Study Author: Natasha Sharma (CC BY 4.0. unless otherwise indicated), University of British Columbia, Bachelors of Environment and Sustainability, 2022*

Landsat 8-7 has regularly acquired images of Beaufort Sea and Mackenzie River Delta through true-color image to show abundance surface melt, fast ice break up, leaf fraction, ice motion, and changes in coastal features during early spring-time. The Beaufort Sea has seen dramatic summer ice losses, particularly in

2009, with regions that were dominated by thick multi-year ice now completely melting out. For comparison, the timeline GIF from 1992 to 2021 depicts different stages of ice breakup in the month of June. Notice the extensive fracturing of Beaufort Sea ice occurring in 1996, but the 2009 and 2014 fracturing appears more widespread until nearly disappearing in 2017.

Similarly, in the high arctic polar desert of Ellesmere Island in Nunavut, Canada, what were once the twin St. Patrick Bay ice caps have now chipped away due to decades of rising temperatures and unusually warm summers. Once the remnants of the Little Ice Age that covered about 7.5 square km and 3 square km across respectively, the formation has reduced to only 5% of their former area and are predicted to extinct within a decade. St.Patrick Bay Ice caps are emblematic of the Arctic change - a reality of how climate change is affecting the whole of Canadian Arctic.

Figure 14.2: Death of the St. Patrick Ice Caps. Sharma, CC-BY-SA-4.0

## 14.5 Pattern Analysis

### 14.5.1 Landscape Pattern Indices

### 14.5.2 Class Pattern Indices

### 14.5.3 Patch Pattern Indices

### 14.5.4 Case Study: Case Study Title Here

Understanding the changes that occur overtime is extremely important for environmental studies. These changes can be monitored using remote sensing data and has been used in a variety of studies. This case study looks at how changes in vegetation health derived from Landsat-8 data were impacted by the burn severity of the 25 569 hectare Little Bobtail Lake wildfire in North-Central British Columbia.

Four vegetation indices were calculated for each of the images in the study. The Normalized Difference Vegetation Index (NDVI) [Rouse et al., 1974], and Tasseled Cap Transformation (TCT) [Crist and Cicone, 1984] were calculated to measure vegetation health. NDVI is calculated as (Equation 1):

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad (1)$$

Additionally, the Normalized Burn Ratio (NBR) [López García and Caselles, 1991], and the Difference Normalized Burn Ratio (dNBR) [Key and Benson,

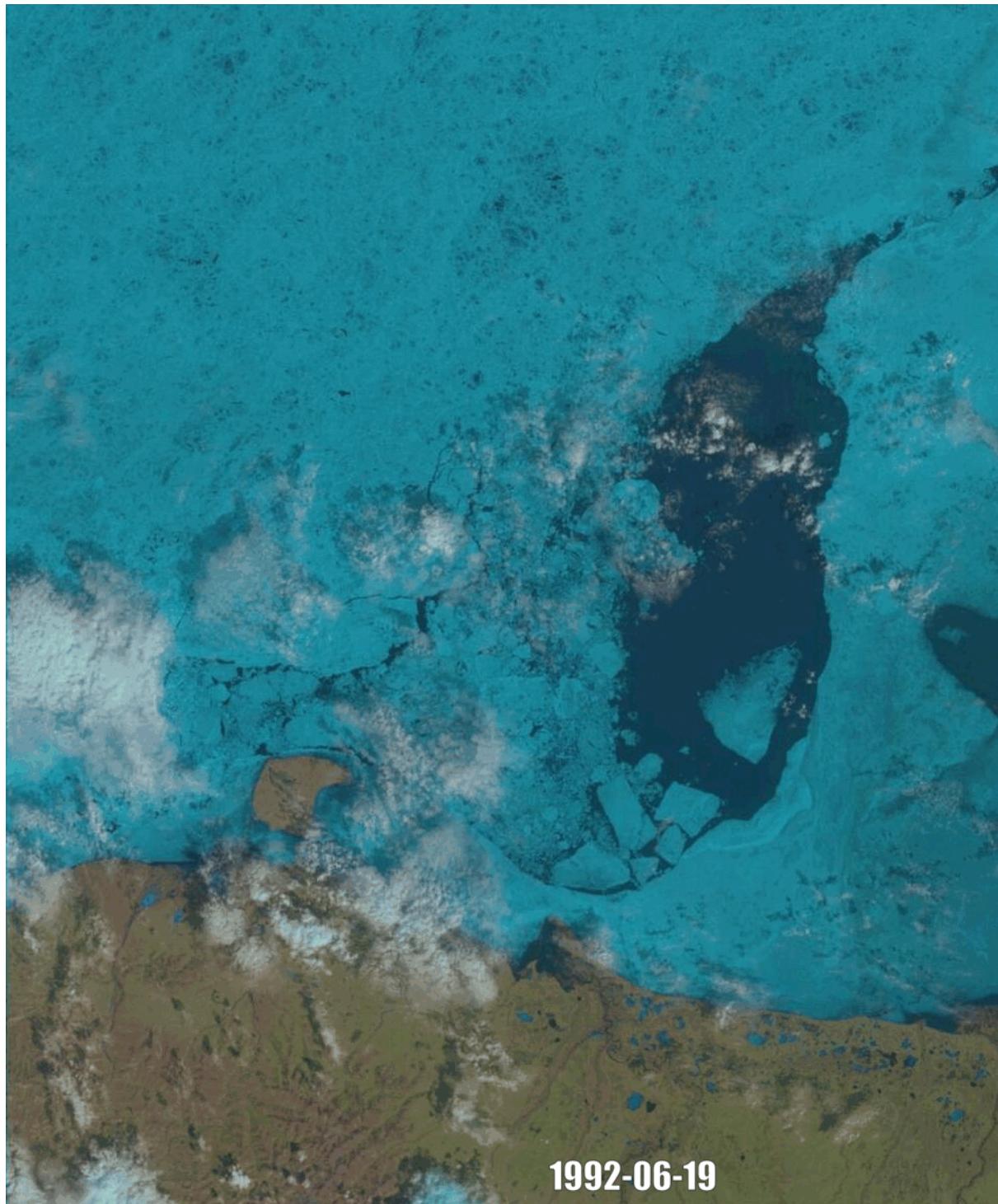


Figure 14.1: True-color image of Beaufort Sea Ice Deterioration. Sharma, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://ubc-geomatics-textbook.github.io/geomatics-textbook/image-analysis.html#fig:14-Beaufort-Sea-Ice>.

Table 14.1: Table 1: Classified Burn Severity Values based on the scaled dNBR values.

dNBR	Classified	Description
< 75	0	Unburned
75-118	1	Low Severity
118-187	2	Moderate Severity
> 187	3	High Severity

2006] were calculated. NBR (Equation 2) and the dNBR (Equation 3) are used to measure burn severity.

$$NBR = \frac{NIR - SWIR}{NIR + SWIR} \quad (2)$$

$$dNBR = NBR_{prefire} - NBR_{postfire} \quad (3)$$

The dNBR values were then scaled for each image to a range of 0-255 by using the following equation (Equation 4):

$$dNBR_{Scaled} = \frac{(dNBR * 1000) + 275}{5} \quad (4)$$

Each of the scaled dNBR images were then classified into four burn severity classes (Table 1). The classified dNBR image for right after the wildfire was polygonised and the burn severity polygons were then used to extract the NDVI and each of the TCT values for each year of the study.

The change in value of the NDVI and the TCT Greenness and Wetness based on burn severity can be seen in the box plots below (Figure 1). This shows the changes in the vegetation index values from before the wildfire and every year after. The burn severity had an impact on these values as there was a greater decrease in value immediately after the wildfire with the higher burn severity as well as a slower return to the pre-fire state. Additionally, based on the vegetation index used, it can be seen that there different rates of return to a pre-fire state. These differences are important to understand when choosing a vegetation index for a study.

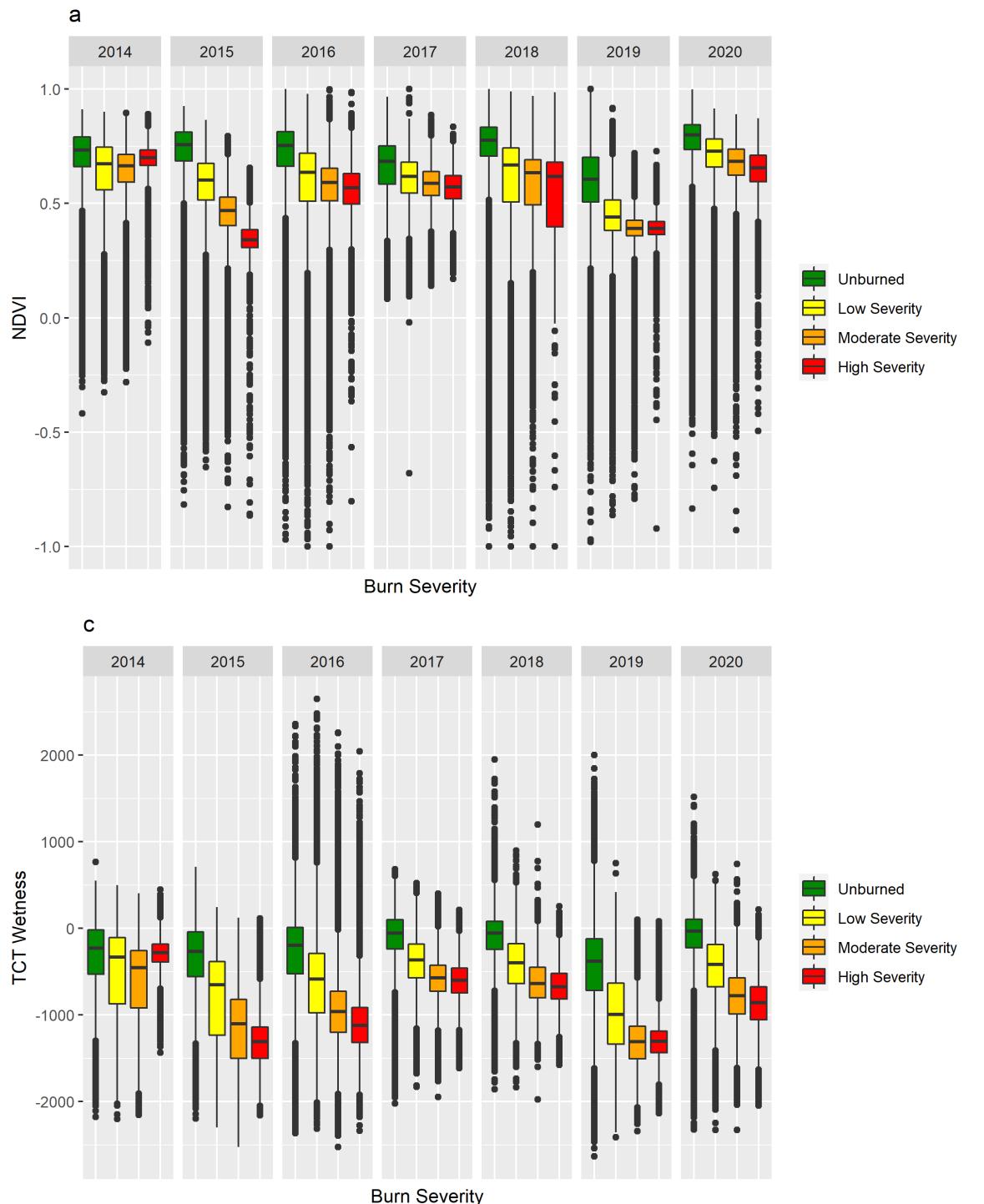


Figure 14.2: Figure 1: Changes in the values of NDVI (a), TCT Greenness (b), and TCT Wetness(c) based on the burn severity of the Little Bobtail Lake wildfire from 2014 (before the wildfire) to 2020 (five years after the wildfire).

Table 14.2: Table 2: Classified Vegetation Health Values based on the NDVI values.

NDVI	Classified	Description
< 0.1	0	No Vegetation
0.1-0.14	1	Sparse Vegetation
0.14-0.5	2	Moderately Healthy Vegetation
> 0.5	3	Healthy Vegetation

Wildfires create spatial patterns on the landscape which is a key factor in forest regrowth. Pattern metrics can be calculated and used to understand the changes in the spatial patterns overtime. Using the `calculate_lsm()` function found in the Landscape Metrics (v1.5.0) R package [Hesselbarth et al., 2019], the Core Area (`lsm_p_core()`) patch metric was calculated for each of the classified dNBR images. Additionally, the NDVI images were classified into four classes (Table 2) and then the Core Area Metric was calculated.

The changes in the Core Area of the burn severity classes showed that the High Severity class had a decrease of 2.6 hectares every year. The Low Severity class had a increase of 1.5 hectares every year and the Moderate and Unburned classes showed little change in Core Area. This was to be expected because as the vegetation regrows in the High Severity areas it changes to a lesser severity (See Map Below). For the vegetation health classes, the Sparse Vegetation class showed a 3.7 hectare increase in Core Area every year, while the other vegetation classes showed little change. These is due to the initial increase in the new vegetation in the first years after the wildfire.

Figure 2:

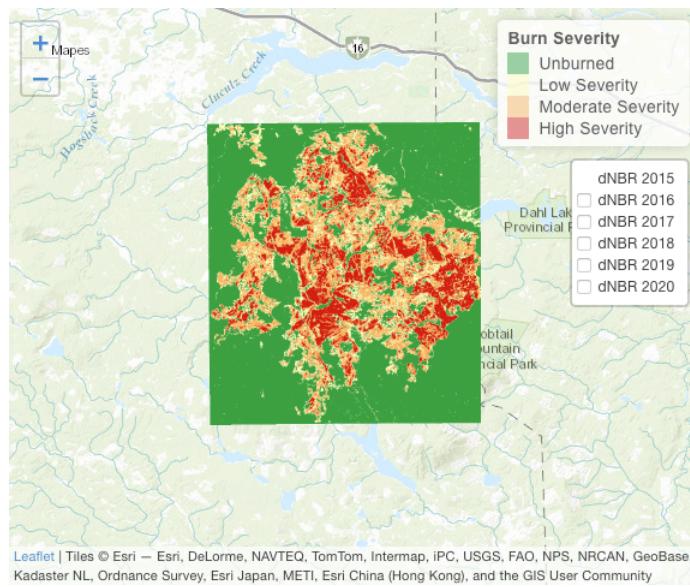


Figure 14.3: Classified Burn Severity images for the Little Bobtail Lake wildfire from 2015 to 2020. Data from Natural Resources Canada [b] and licensed under the Open Government Licence - Canada. Pickell, CC-BY-SA-4.0. Animated figure can be viewed in the web browser version of the textbook: <https://ubc-geomatics-textbook.github.io/geomatics-textbook/image-analysis.html#fig:14-dnbr-leaflet>.

## **Chapter 15**

# **LiDAR Acquisition and Analysis**

Written by Francois du Toit and Paul Pickell

Remotely sensed data has traditionally relied on sensors to passively collect reflected energy from the Earth's surface. As you have read in the preceding chapters, the information that we can derive from these sensors is immense; from landscape level satellite analyses to individual tree vigor assessments. One aspect that is more difficult to characterize however, is what a landscape looks like in three dimensions. LiDAR is changing how we can interpret landscapes and forests by producing its own source of energy; this allows us to ask questions not only about the top of the Earth's surface, but also about the structure of the forest, and what the ground looks like beneath it. In this chapter, we discuss what LiDAR is, as well as how we can use it in multiple different contexts.

### **Learning Objectives**

1. Understand what LiDAR is and how it works
2. Understand what we can do with LiDAR data, and what products we can generate
3. Understand the basic processing steps required to use LiDAR data for forestry and ecological analysis

### **Key Terms**

Light Detection and Ranging (LiDAR), Laser, Global Navigation Satellite System (GNSS), Inertial Measurement Unit (IMU), Discrete Return, Full Waveform, Surface Models, Area-Based Approach (ABA), Individual Tree Crown

Detection (ITD), LiDAR Metrics, Tree Segmentation.

## 15.1 What is LiDAR?

**LiDAR** stands for **Light Detection And Ranging** (sometimes written as lidar, or LIDAR), and is an active remote sensing technology. A laser scanner and time of flight principles are used to collect three dimensional (3D) data. LiDAR systems are made up of three components; a laser-scanning device, an accurate global navigation satellite system (GNSS), and an inertial measurement unit (IMU). The laser-scanning device emits pulses of light and measures the time it takes for energy to be reflected to the device. The GNSS receiver allows the position of the laser to be determined in space, while the IMU records the orientation of the laser (i.e. roll, pitch, and yaw [White et al., 2013]). Figure 15.1 illustrates all of the necessary components of a LiDAR system. When discussing LiDAR in an airborne context (i.e. the unit is being flown), we can call it airborne laser scanning (ALS), and if the unit is in a fixed position on the ground it is called terrestrial laser scanning (TLS).

## 15.2 How Does LiDAR Work?

Typical LiDAR systems use the time of flight method to produce 3D data. A laser ranging instrument produces a short, intense pulse of light from the instrument to a target being measured. Some of this energy is then reflected back to the instrument, where it is recorded (as seen in Figure 15.2). Since the speed of light, and the location of the laser ranging instrument is known, we can calculate the position of the target by timing how long it takes between the the pulse being emitted and received. If we shoot many pulses of light towards a target, we can create a 3D point cloud of our target (see Figure 15.3 for an example of a forest scene). Modern LiDAR systems can emit hundreds of thousands of pulses per second, which means LiDAR point clouds can contain millions of points, and be several gigabytes large.

Since a LiDAR point cloud is a file containing the 3D location of points representing objects on the Earth's surface, there are several parameters to take note of. It is important to know what the technical specifications of the data collection parameters were (for example, how many points per square meter do we have?), and how the data was collected (what platform was used?). A LiDAR dataset includes several pieces of information for each 3D point.  $X$ ,  $Y$ ,  $Z$  location data tells us where each point is (usually to the centimeter scale), and a GPS time stamp is included for each point (*gpstime*). Additional information such as *return number*, *scan angle*, *classification*, and *intensity* are also included with the file, which will be discussed in more detail below (Components of a LiDAR System).

The most common file format for LiDAR files is called the LAS file format

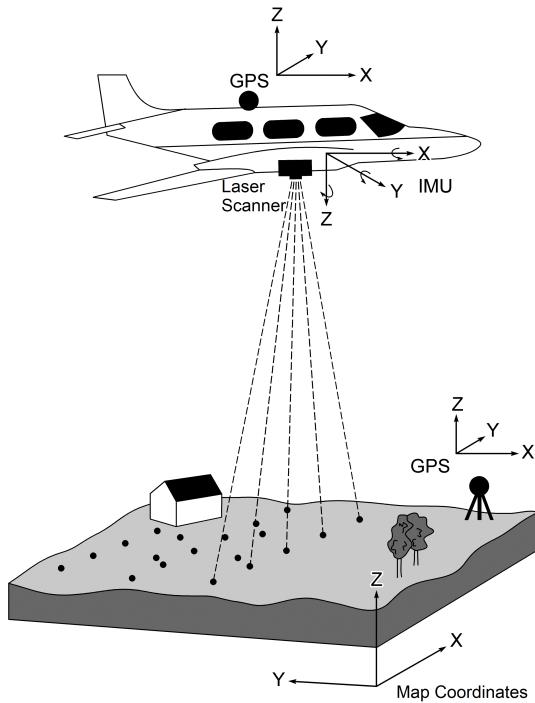


Figure 15.1: Overview of the components of a LiDAR System. As is common for large scale data acquisition, the laser scanning is placed on board an aircraft and scans the Earth below it. The aircraft has on board GPS/GNSS, as well as an inertial measurement unit (IMU). A GPS base station can be used to post-process data and increase spatial accuracy. ('LiDAR-i lend', Marek9134 [2012], CC-BY-SA-3.0).

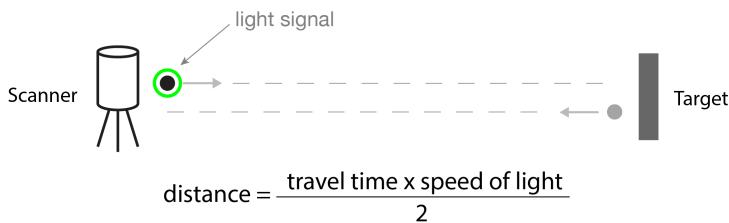


Figure 15.2: Concept of LiDAR. A light signal is emitted by the scanner and reflected off the target. ('Concept of LiDAR', Cartographer3d [2021], CC-BY-SA-4.0).

(.las). This file format was originally designed for 3D point cloud data, and is a free alternative to proprietary systems or a generic ASCII file interchange system. The main benefits of this file format are that it is relatively quick, can be used by any system, and stores information specific to the nature of LiDAR data without being overly complex [American Society for Photogrammetry & Remote Sensing, 2019]. More information regarding the file type specifications is available from the ASPRS's webpage LASer (LAS) File Format Exchange Activities.

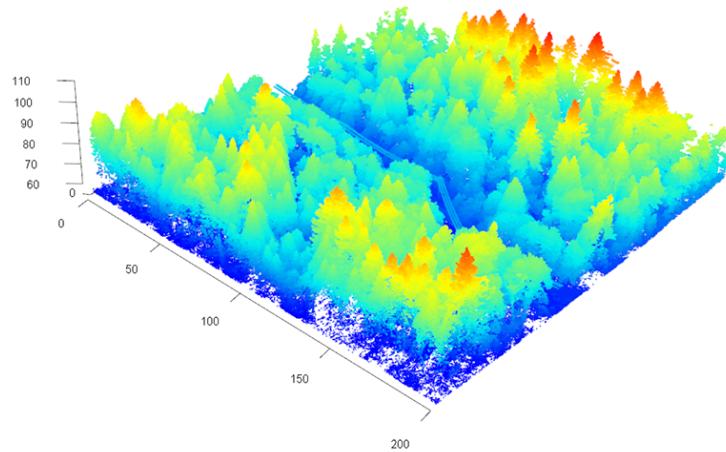


Figure 15.3: (ref:las-denoise-caption)

### 15.3 LiDAR History and Use

LiDAR for forestry and ecology has been used to monitor vegetation structure, stream properties, and topography among other uses. The earliest versions of LiDAR were known as profiling systems and first investigated in the 1960's [Nelson, 2013]. A major limitation of early LiDAR profiling systems were that the system was locked in a near-nadir position (i.e. along-track path), which meant that only transects could be collected, as opposed to a larger swath width as with modern systems [Nelson, 2013, Lim et al., 2003]. Technological advancements meant that early scanning LiDAR systems became more common in the 1980s, although point densities (measured as points per square meter,  $\text{points} \cdot \text{m}^{-2}$ ) were low. Densities of  $1\text{-}5 \text{ points} \cdot \text{m}^{-2}$  limited researchers to area based measurements of forest volume and biomass [Nelson, 2013]. These limitations were primarily due to the LiDAR sensor, as well as data storage difficulties. Technological improvements such as increased pulse rates and smaller footprints (leading to increased point density), and increased storage capacity have allowed researchers to look at individual trees as well as whole forests [Jakubowski et al.,

2013]. Previously limited to static LiDAR sensors, ALS point cloud densities of  $1000 \text{ points} \cdot \text{m}^{-2}$  are now entirely feasible. Figure 15.4 shows two different point clouds densities, and how difficult it would be to delineate individual tree crowns at low point densities.

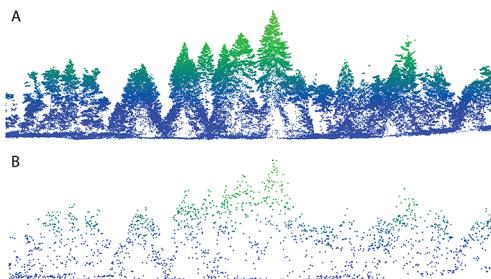


Figure 15.4: A cross section of a high density point cloud (A,  $80 \text{ points} \cdot \text{m}^{-2}$ ), and a lower density point cloud (B,  $1 \text{ point} \cdot \text{m}^{-2}$ ). (Du Toit, CC-BY-4.0).

Unlike passive remote sensing technologies, LiDAR has the advantage of being able penetrate forest canopies; this means that we are able to detect the ground, and also characterize both vertical and horizontal vegetation structure. Since the data is extremely accurate, surface models developed using LiDAR are used in many fields where human-made and natural environments need to be mapped. LiDAR derivatives such as surface models are used in hazard assessment (see here: Jaboyedoff et al. [2012]), forestry (read more here: Goodbody et al. [2021]), wet area mapping (here: Eash et al. [2018] and Zurqani et al. [2020]), geologic/geomorphological mapping, and agriculture. In this chapter, we will discuss the use of LiDAR primarily in a forestry context, where raw data and derivatives are used for volume and biomass estimation, as well as individual tree crown analyses.

## 15.4 Components of a LiDAR System

### 15.5 Lasers

**Lasers** are a very important component of LiDAR systems. Here we will discuss some basic concepts, as well as the technical parameters that are important when interpreting LiDAR data. LiDAR lasers are typically beams of **near-infrared** light ( $800 - 1,550 \text{ nm}$ , typically  $1,064 \text{ nm}$ ), and are used because these wavelengths are considered eye safe [White et al., 2013]. Green wavelengths are used for bathymetric LiDAR (i.e. water penetrating LiDAR), but are less common, and not typically used on land [UF Geomatics - Fort Lauderdale, 2016c]. Laser scanners use rotating or oscillating mirrors in order to ‘scan’ a scene in multiple dimensions [UF Geomatics - Fort Lauderdale, 2016b]. When these scanners are placed on a moving platform (e.g. a plane), we can cover

large areas [UF Geomatics - Fort Lauderdale, 2016b]. LiDAR sensors scan in a variety of ways; zig-zag, rotation mirror line, and push broom scanners are the most common, and use the same principles as those described for sensors in **Chapter 12 (double check, dynamically link)** [UF Geomatics - Fort Lauderdale, 2016b].

**Beam divergence** is a property that refers to how wide the light beam becomes when it intercepts an object, and can be used to differentiate LiDAR instruments. Small-footprint LiDAR describes beam diameters intercepting the surface at < 1 m, while large-footprint intercepts the surface at around 5 - 25 m [Lim et al., 2003]. Small-footprint LiDAR is primarily what is used in forest inventory and ecology studies, and has high accuracy, as well as the ability to produce high sampling densities. For these studies, beam divergence is typically 0.15 – 2.0 mrad [White et al., 2013]. The concept is illustrated in Figure 15.5, and the equation is shown below in (15.1), where  $D$  is the footprint diameter,  $h$  is the flight height (m), and  $\gamma$  is the laser beam divergence (mrad).

$$D = h\gamma \quad (15.1)$$

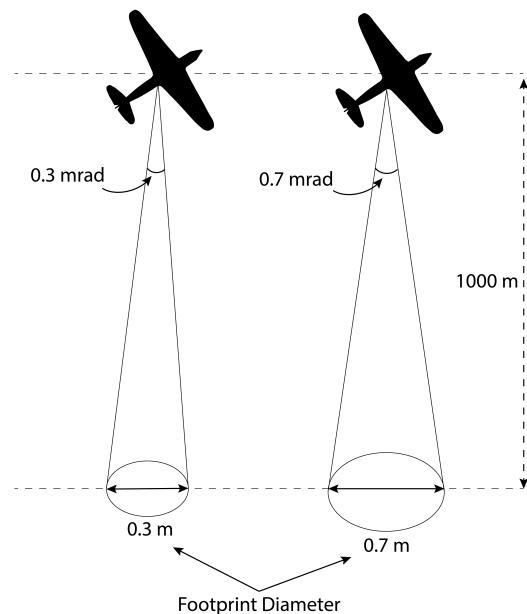


Figure 15.5: Two examples of how beam divergence affects footprint size. (Du Toit, CC-BY-4.0).

The amount of energy that is reflected off an object and back to the sensor is known as **intensity**. Target reflectivity is not directly related to the LiDAR laser itself, but influences whether the return has enough intensity to register

with the LiDAR sensor. In addition to surfaces having different properties, the angle of incidence of the laser also affects how much energy the sensor receives. This ‘field of view’ is known as **scan angle** and can be customized, but lower scan angles are generally preferred (<25°, White et al. [2013]). The concept of scan angle is the same as discussed in Chapter 12 and illustrated in Figure 15.6. Commercial LiDAR units on airplanes typically have stronger lasers than drone mounted or mobile laser scanners, which allows them to fly higher and cover more area.

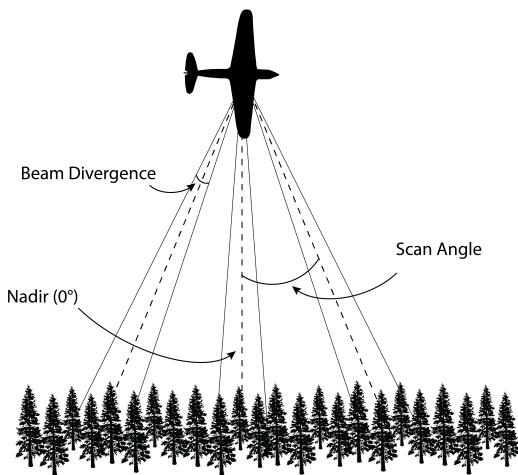


Figure 15.6: Scan angle is the angle from nadir at which the laser is pointing. The scan angle and aircraft flight height together are responsible for swath width. (Du Toit, CC-BY-4.0).

All LiDAR sensors emit pulses at a certain rate (**pulse rate**), which can be given as pulses/second, or hertz (Hz). Pulse rate is highly variable (and often programmable on sensors), and along with **scan rate** (the number of scan lines per second) and **flight speed** is responsible for what density a point cloud can be. Pulse rates are commonly 50,000 - 200,000 Hz [White et al., 2013]. All of this information means that each LiDAR data acquisition campaign has the potential to be rather unique, and the parameters italicized above need to be taken into account when doing analysis. All of these factors together affect how an individual pulse interacts with the target.

Several technical aspects of lasers are recorded by the instrument (and usually provided as a flight summary or flight specifications by vendors), while some of it can also be found in the `1as` file. We can usually find information that impacts the entire flight in the flight specifications, such as flight height, scan rate, and beam divergence. These aspects shouldn’t change for the entire duration of the flight. In contrast, information located in the `1as` file affects individual points. Here we can find information such as scan angle, return number, and intensity.



Figure 15.7: An example of a LiDAR Unit. ('LiDAR machine', The Center for International Forestry Research (CIFOR) [2014], CC-BY-NC-ND-2.0).

## 15.6 Position and Orientation

## 15.7 Global Navigation Satellite Systems

In order for us to use the calculation from Figure 15.2, we need to know the exact position of the scanner in space. For airborne and mobile platforms, we can do this by using the **global navigation satellite system (GNSS)**, as well as post processing using a local GNSS reference station. This helps us to improve accuracy by not comparing our position to a known local position in order to ensure our location is known as accurately as possible. GNSS concepts are covered further in Chapter 4.

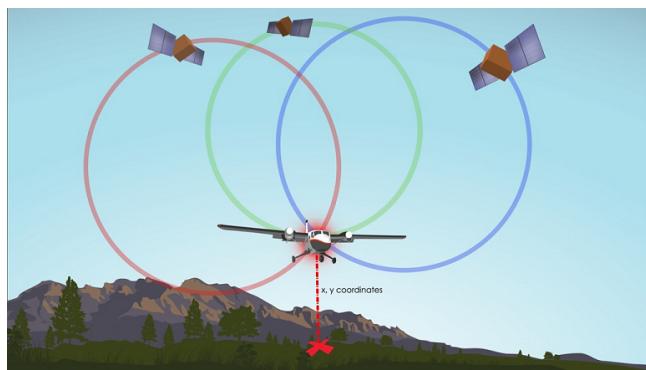


Figure 15.8: An on board GNSS is required to provide an accurate location of the LiDAR scanner in space. ('A key component of a lidar system is a GPS', NEON Education [2014], CC-BY-NC-SA-2.0).

## 15.8 Inertial Measurement Unit (IMU)

An **inertial measurement unit (IMU)** consists of gyroscopes and accelerometers and measures the attitude and acceleration of the aircraft along the X, Y, Z axis. This data is combined with the GNSS data to provide a precise location of the scanner in space. This information becomes very important for airborne platforms where wind conditions can cause the orientation of the scanner to change subtly. We refer to the orientation of the platform in space by using the terms pitch, roll, and yaw, to describe which way the scanner is facing (see 15.1).

## 15.9 Clocks

LiDAR point data needs to be synced with positioning data in order to know exactly where the point is in space. To do this, a very accurate GNSS clock is used to time stamp the laser scanning data [UF Geomatics - Fort Lauderdale, 2016b]. Accurate clocks are imperative for producing accurate point clouds; one nanosecond (i.e., one billionth of a second) corresponds to a 30 cm travel distance, as seen in Figure 15.11 [UF Geomatics - Fort Lauderdale, 2016c]!

## 15.10 Platform

LiDAR units can be attached to a variety of platforms. Traditionally, LiDAR units for forestry research were mounted on airplanes and helicopters, as units were large and cumbersome, however ground based units such as terrestrial laser scanning (TLS), and mobile laser scanning (MLS) have also been developed. These units tend to have a very high point density, and TLS is often used in modeling tree architecture.

## 15.11 Airplanes and Helicopters

Airplanes and helicopters are still the most common platforms for LiDAR data collection. This is due to their ability to collect large amounts of data in one acquisition. LiDAR units that are designed for airplanes are more powerful than those designed for drones, leading to increased penetration in forests, even though the flight height is significantly increased. Data storage issues can also be mitigated as the platform is not as sensitive to weight restrictions when compared to drones. With increased pulse rates, data acquisitions can also be dense enough to do individual tree crown work. Since helicopters are capable of flying at lower altitudes, slower speeds, and following terrain, they are capable of collecting more dense data, but higher operating costs [White et al., 2013]. Typical point densities range from 5 - 200 points · m<sup>-2</sup>.

## 15.12 Drones

As laser units have decreased in size, we have been able to mount LiDAR units to smaller platforms. Drones (also known as unmanned aerial vehicles (UAV) or remotely piloted aerial systems (RPAS)) are extremely convenient to collect high density point clouds over small areas. They typically fly relatively close to the forest (e.g. 50 m above the trees) and fly a pre-defined flight route so ensure evenly spaced data collection. Typical point densities range from hundreds to thousands of points per meter squared. A major limitation of drones is currently battery life, as acquisition campaigns are limited by how long the drone can fly for (approximately 20 - 30 minutes). This means that collecting large amounts of data is difficult in remote locations.



Figure 15.9: Drone mounted LiDAR. This technology is rapidly developing, with many companies working towards creating units capable of acquiring high density datasets over large areas. ('LiDARUSA Snoopy 120 LiDAR', Mc Clurhans [2019], CC-BY-SA-4.0).

## 15.13 Mobile Laser Scanning

Mobile Laser Scanning (MLS) includes people (sometimes called 'backpack LiDAR') and moving vehicles. This version of LiDAR has been used by transportation and engineering companies to precisely map things like road or building conditions, and as the units have reduced in size there has been an increase in interest for use in forestry. Current MLS units such as panel 2 in Figure 15.10 can be carried around by the user in a forest easily. The point density for that specific unit is approximately  $6,000 \text{ points} \cdot \text{m}^{-2}$  while the battery can last for over an hour. Typical point densities are highly dependent on what the MLS is mounted to; and can range from hundreds to thousands of points per square meter.

### Terrestrial Laser Scanning

Terrestrial Laser Scanning (TLS) is a static system, meaning that it is placed in one location, takes measurements of a scene, and can then be moved (see panel 1 of Figure 15.10). These units are used by engineers for surveying, and are used by foresters to precisely measure trees. This information can be used to build quantitative structure models (QSM), to precisely estimate tree parameters. Unlike the other platforms mentioned here, TLS is not necessarily a fast data collection technique. Due to occlusion, the TLS instrument needs to be placed in multiple positions around a plot to be able to fully visualize the trees/plot, which means that a one hectare plot can take three to six days to survey [Wilkes et al., 2015]! Typical point densities are in the thousands of points per meter squared.

## 15.14 Satellite

Satellites are a relatively new platform for LiDAR sensors. Two current satellite sensors are highlighted here; ICESat-2 (Ice, Cloud and land Elevation Satellite) and GEDI (Global Ecosystem Dynamics Investigation). Both are examples of large-footprint, profiling LiDAR systems. ICESat-2 was launched in 2018, with the objectives to determine sea-ice thickness and measure vegetation canopy height for large scale biomass estimation [NASA and Neumann, 2021]. The ICESat-2 LiDAR sensor is composed of 6 beams in 3 pairs, each with a footprint size of 13 m [NASA and Neumann, 2021]. GEDI is a LiDAR sensor that was installed on the International Space Station in 2018, with 3 lasers producing 10 parallel tracks of observations, and a footprint size of 25 m [NASA and University of Maryland, 2021]. The GEDI sensor can be seen in panel 4 of Figure 15.10.

## 15.15 Types of LiDAR

### 15.16 Discrete Return

**Discrete return** LiDAR is the most common type of LiDAR. Here, the reflected energy from a pulse is converted into return targets referenced in time and space [White et al., 2013]. Discrete return systems can often record multiple returns (often 5 or more), as long as the energy from each return meets the energy threshold to be classified as a return. Returns are classified per pulse, by the arrival sequence in which the return signals are detected by the sensor [UF Geomatics - Fort Lauderdale, 2016a]. In the case where a laser pulse intercepts a solid object (such as a building), only a single return would occur; in contrast if the laser pulse can penetrate through the object (e.g. the forest), some of the energy will be returned to the instrument where the pulse intercepts stems/branches/leaves [White et al., 2013]. Figure 15.11 shows an example of this; a discrete return system would interpret the returned echo as two returns,



Figure 15.10: LiDAR Platforms 1: A terrestrial laser scanner (TLS). ('Lidar P1270901.png', Monniaux [2007], CC-BY-SA-3.0). 2: A hand-held mobile laser scanner (MLS). (Du Toit, CC-BY-4.0). 3: Car mounted MLS. ('3D mobile mapping unit', Oregon Department of Transportation [2016], CC-BY-2.0). 4: GEDI, a spaceborne LiDAR sensor. ('A New Hope: GEDI to Yield 3D Forest Carbon Map', NEON Education [2014], CC-BY-2.0).

as there are two peaks in the returned echo corresponding to a canopy and a ground hit. In a forested landscape, first returns often represent the upper canopy, while last returns correspond to the ground or objects near the ground [White et al., 2013].

## 15.17 Full Waveform

**Full waveform** LiDAR systems record reflected energy from each laser pulse as a continuous signal [White et al., 2013]. Traditionally, discrete return systems would only digitally store the discrete returns, without recording the complete analog return [UF Geomatics - Fort Lauderdale, 2016a]. In contrast, full waveform systems digitize the entire analog signal at high sampling rates [UF Geomatics - Fort Lauderdale, 2016a]. This can be seen on the right hand side of Figure 15.11. This type of LiDAR creates much larger file sizes which were an issue in the past, but with modern storage capabilities it is becoming a viable alternative to discrete return LiDAR.

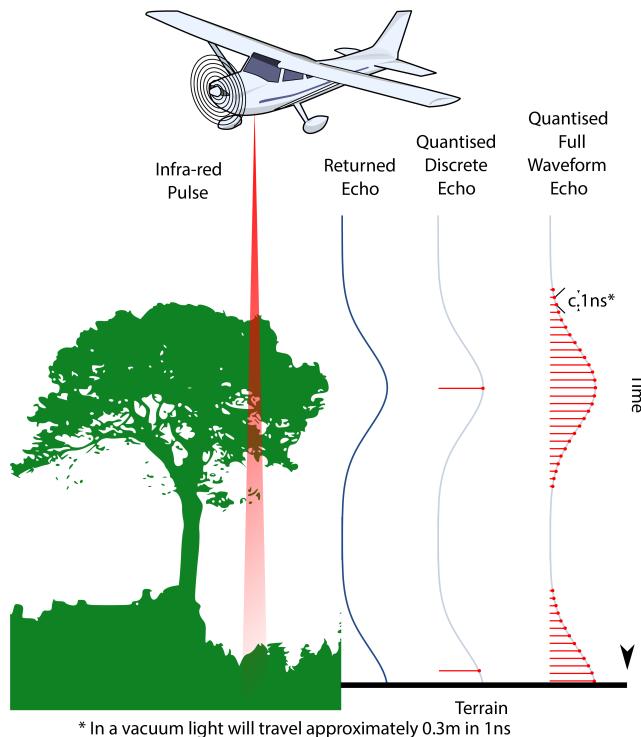


Figure 15.11: Difference in the response between discrete echo and full waveform LiDAR. ‘Airborne Laser Scanning Discrete Echo and Full Waveform signal comparison’, Beck [2012b], CC-BY-SA-3.0).

## 15.18 Emerging Technology

**Single Photon LiDAR** is an emerging technology that allows wall-to-wall mapping detecting photons reflected off of surfaces more efficiently than traditional LiDAR sensors [Swatantran et al., 2016]. While traditional LiDAR systems emit high energy laser beams, single photon systems transmit shorter, lower energy pulses [Swatantran et al., 2016]. These systems use a green laser (532 nm), and due to the high efficiency nature of the technology can acquire high density point clouds ( $12 - 30 \text{ points} \cdot \text{m}^{-2}$ ) up to 30 times faster than traditional systems, and operate at higher altitudes [Swatantran et al., 2016]. The system is also capable of penetration semi-porous obscurations such as vegetation, ground fog and thin clouds during daytime and night-time operations [Swatantran et al., 2016].

**Multispectral LiDAR** units such as the Optech Titan sensor operate three distinct wavelengths of 1,550 nm, 1,064 nm, and 532 nm [Morsy et al., 2017]. This technology enables 3D spectral information to be captured; for example, NDVI can be calculated for the 3D point cloud. Combining multispectral LiDAR at three wavelengths allows for higher reliability and accuracy compared to single-wavelength LiDAR [Morsy et al., 2017], however this technology is not yet widespread.

## Call Out

Specifications for LiDAR data acquisitions are highly variable. It is important to know what type of LiDAR you are using, and to understand the technical specifications of the point cloud before diving into any analysis.

## 15.19 LiDAR Derivatives and Analysis

LiDAR point clouds require post-processing to be useful. Often, point clouds contain noise that needs to be removed, while the points need to be classified to produce derivatives. Once a point cloud has been cleaned and classified, we can create a variety of products to describe the Earth's surface, as well as describe vertical characteristics of our point clouds. The most important points to identify in a point cloud are ground and non-ground returns; these points are crucial for deriving **surface models**, such as: digital elevation models, digital surface models, and canopy height models (discussed in more detail in Chapter 9).

## 15.20 Bare Earth Elevation

As we learned from Chapter 9, Digital Elevation Models (DEM) are raster products that represent elevation of the Earth's surface above a reference vertical datum such as sea level [White et al., 2013]. Using a classified point cloud, we

can isolate the points that are categorized as ground (`Classification == 2`) and create a raster that represents the ground (see Figure 15.12 below). DEMs are also called digital terrain models (DTMs), although a DTM can include vector features that represent natural features such as river channels, cliffs, ridges, and peaks. DEMs have a variety of uses, from terrain modeling for watersheds, to road design [Roussel et al., 2021]. In addition to being used on their own, DEMs can be used to normalize point clouds; this is the process of subtracting the ground elevation from a point cloud in order to place the points on a flat plane, so that each point represents the height above the ground, not absolute height. To create a DEM, various interpolation techniques are available - these techniques are covered in more detail in Chapter 10, but common algorithms include inverse distance weighting (IDW), kriging, and k-nearest neighbour (KNN).

When creating a DEM from a LiDAR point cloud, there are a few important considerations to make. The first is what resolution the DEM should be created with. The density of the point cloud is what dictates this; if the density of a point cloud is low, there is a lower likelihood of a point representing the ground. This means that a lower resolution DEM should be created. Conversely, high density point clouds can produce higher density DEMs. Additionally, forest/vegetation type is important. Dense, coastal forests in BC are likely to have high canopy cover (and therefore fewer ground points) than the more sparse subalpine forests in the province's interior.

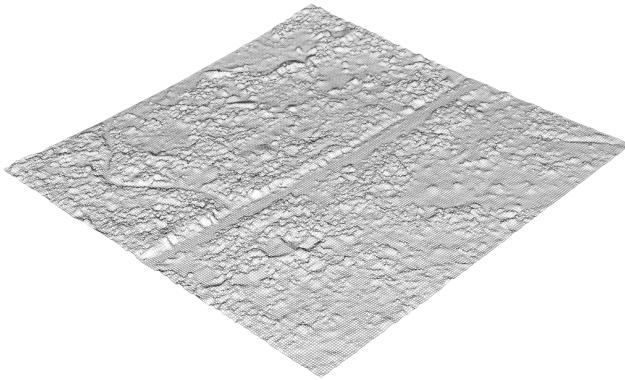


Figure 15.12: A digital elevation model (DEM) produced from LiDAR.(Du Toit, CC-BY-4.0).

## 15.21 Digital Surface Model and Canopy Height Models

Digital surface models (DSMs) represent the height of surface features like trees and buildings above the ground. The ground is defined as the elevation of the

terrain above a reference vertical datum. Thus, a DEM is used to reference heights represented in a DSM. In contrast to a DEM, the DSM captures the natural and built features of the environment, and can be thought of as a table cloth placed over a scene. When a point cloud is normalized by subtracting a DEM from a DSM, the derived surface can be called a canopy height model (CHM, Figure 15.13), and represents the height of the canopy above ground level [White et al., 2013]. While the elevation values are different, both surfaces can be derived using the same algorithms [Roussel et al., 2021].

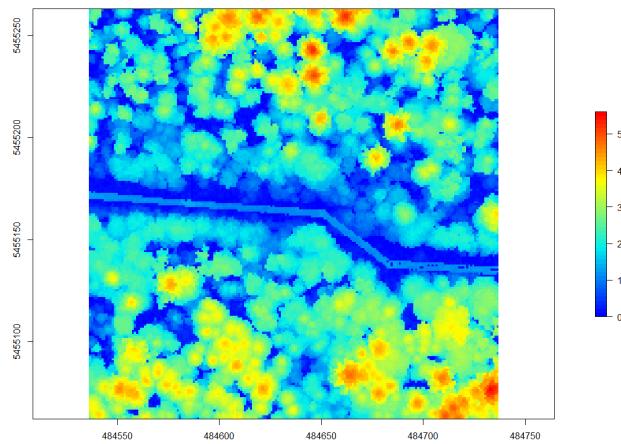


Figure 15.13: A canopy height model (CHM), also known as a digital surface model (DSM) produced from LiDAR. (Du Toit, CC-BY-4.0).

## 15.22 Area Based Approach vs. Individual Tree Crown Approach

LiDAR analysis in forestry uses two broad approaches, an **area-based approach (ABA)**, or an **individual tree crown detection approach (ITD)**. The ITD approach locates individual trees and allows the estimation of individual tree heights and crown area. From this, other metrics such as stem diameter, number of stems, basal area and stem volume can be derived [Hyppä and Inkinnen, 1999]. In contrast, the main goal of an ABA is to generate wall-to-wall estimates of inventory attributes, such as mean height, dominant height, mean diameter, stem number, basal area, and volume of the stands [Næsset, 2002]. Since the ABA produces wall-to-wall estimates, the products are usually in a raster format. The ABA approach is what is most often used for forestry inventories, and is an extensive best practices guide for Canada was produced in 2013 by White et al. [2013]. Examples such as Tompalski et al. [2019] show the power that using an ABA can have.

When doing either of these analyses, we typically produce **LiDAR metrics**. These metrics are descriptive statistics of the point cloud over a unit area; for example Figure 15.14 shows the maximum height of 10 x 10 m cells. How we define that unit decides what kind of analysis we are able to do, and what kind of inferences are possible (i.e., ABA vs ITD). An ITD approach has a ‘unit’ size of one tree, whereas the ABA typically uses a grid size of 20 x 20 m. Both approaches require ground truthing, and for an ABA approach, plot sizes are typically designed to be of a similar area to the chosen grid size [White et al., 2013].

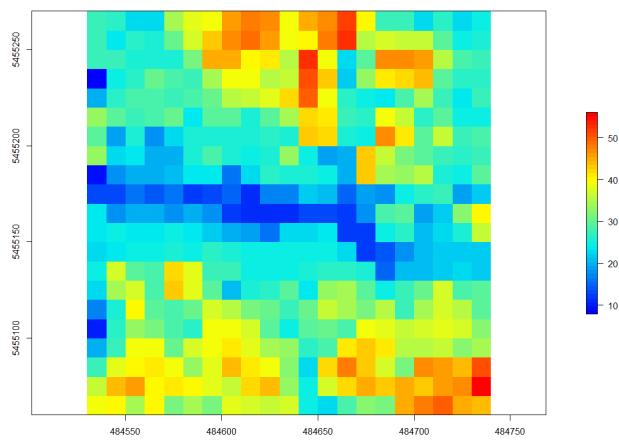


Figure 15.14: Maximum heights of 10 meter cells using an area-based approach. (Du Toit, CC-BY-4.0).

## 15.23 Tree Segmentation

**Tree segmentation** is required when undertaking an ITD approach. Increased point densities as well as other advancements in LiDAR systems mean that this approach is becoming operationally relevant. Tree segmentation is an attempt to extract individual trees from a point cloud. Since point clouds do not include information regarding what point belongs to which tree, we need to classify the points in a similar way to how we would classify points that represent the ground. Several different algorithms exist in order to detect individual trees and then segments them. Most algorithms follow the logic that tree tops will be near the top of the point cloud, and once the tree top has been identified, a region growing algorithm can be used to ‘grow’ the point cloud downwards.

Once we have segmented a point cloud, we can produce metrics for individual trees. These metrics can be used to help to identify structural traits of trees, which can then be used to identify different tree species, or to characterize different forest types. Ground based LiDAR scanners (MLS and TLS) are also

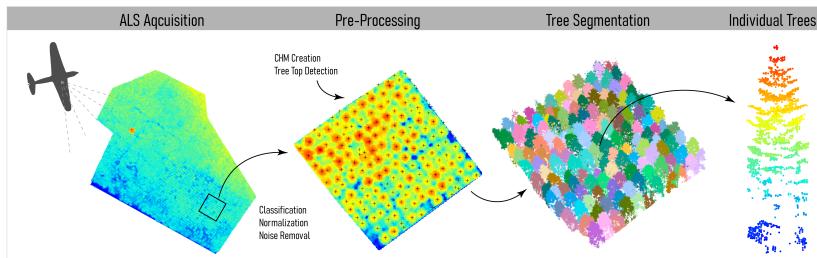


Figure 15.15: A typical LiDAR processing workflow. After acquiring data, pre-processing is necessary to clean and normalize the point cloud. After this a CHM can be created to detect tree tops, before segmentating the point cloud based on these points. (Du Toit, CC-BY-4.0).

used to very accurately estimate tree, trunk, and branch volumes.

## 15.24 Sources of Error

As with other data collection techniques, there are a few sources of error to be aware of. Accuracy of the components are important; if the GNSS is accurate to 1 m, it will cause significant issues compared to a GNSS that is accurate to 2 cm! Additionally, if the IMU has errors, we can have trouble locating ourselves in space, which affects where the laser return is ‘placed’ in 3D space. Typically, we are supplied with estimates of accuracy by the data supplier.

Aside from positional errors, there can also be issues with how we correct for atmospheric conditions, and how it absorbs light. The target surface should also be taken into account, as it can cause multipath errors (where the pulse is reflected off multiple surfaces before returning to the sensor, see Figure 15.17), and occlusion (lasers can not penetrate through solid objects). Finally, areas with highly variable terrain can lead to uncertainty in position, especially when with low density LiDAR.

## 15.25 Software and Analysis Tools

As LiDAR acquisition becomes cheaper, more tools are becoming available to do analysis. In this chapter, we use the a free and open source R package liR; which can be used for the entire process of analyzing a point cloud. Several other options exist, such as A Shiny-based Application for Extracting Forest Information from LiDAR data (also in R), the Digital-Forestry-Toolbox which is available for MATLAB, the software FUSION, which has been developed by the USDA Forest Service, and finally the free and open source GIS software QGIS. Paid software is also frequently used in LiDAR processing (examples include Esri’s ArcGIS Pro, and LAStools), although the cost can be prohibitive.

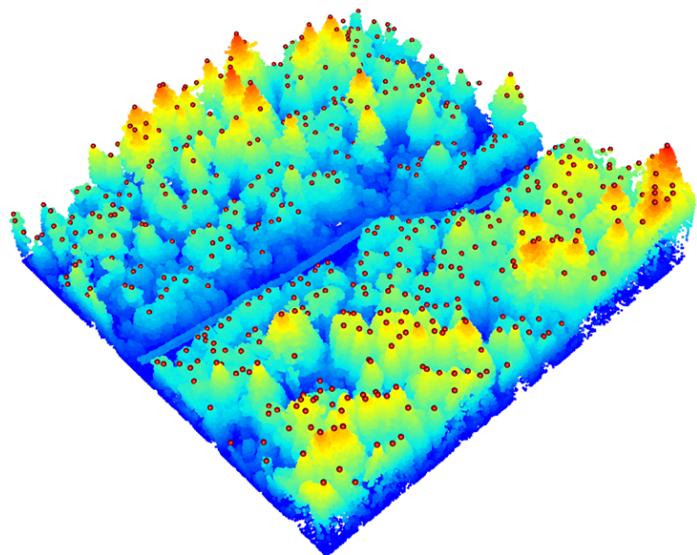


Figure 15.16: Example of tree tops detected using the `find_trees` function.  
(Du Toit, CC-BY-4.0).

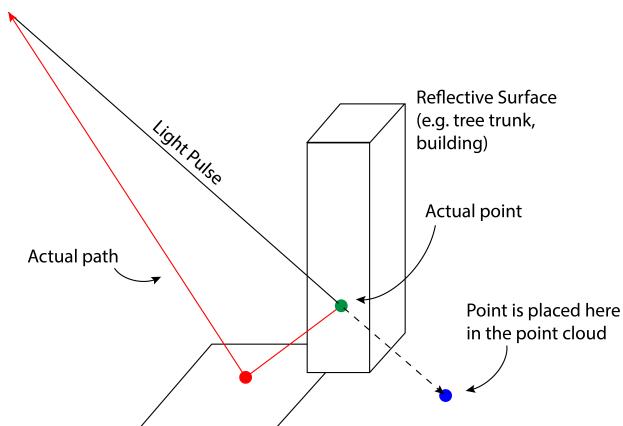


Figure 15.17: Example of multipath errors that can occur. The green dot is where our point is actually located in space, while the blue point shows where the point is placed in the point cloud. The red line shows the path of the reflected pulse. (Du Toit, CC-BY-4.0).

Finally, the open source software CloudCompare can be incredibly useful for both viewing and manually clipping/editing point clouds.

## 15.26 Case Study: Creating LiDAR Metrics from a Raw Point Cloud

For this case study, we will be using a clipped .las file from the 2018 open LiDAR dataset of the City of Vancouver and UBC Endowment Lands in British Columbia (we randomly selected a 200 x 200 m portion of the ‘4840E\_54550N’ tile using CloudCompare), and the `lidR` package in R [City of Vancouver, a],[Roussel et al., 2020]. The script to process this data is included in the data folder of the GitHub repository, and you can use the online `lidR` book to get a more in depth understanding of the functions we apply below [Roussel et al., 2021].

The first step when looking at LiDAR data is to inspect it; we recommend using the free software CloudCompare, or plotting the .las file in the `lidR` package. Once we have a sense of our data, we can clean and filter the data to remove noise. Our cleaned dataset can then be used to create a DEM; first we need to classify ground points, followed by using an algorithm to rasterize our new ground points to create a surface. This is an essential step that could require quite a bit of tweaking depending on what you want to use the DEM for. It is important to take density into account; a sparse point cloud will not be good for a high resolution DEM! In our case, the DEM is used to *normalize* the point cloud. Normalization removes the effect of terrain on above ground measurements, allowing comparisons of vegetation heights as you can see in the very uniform point cloud in Figure 15.18.

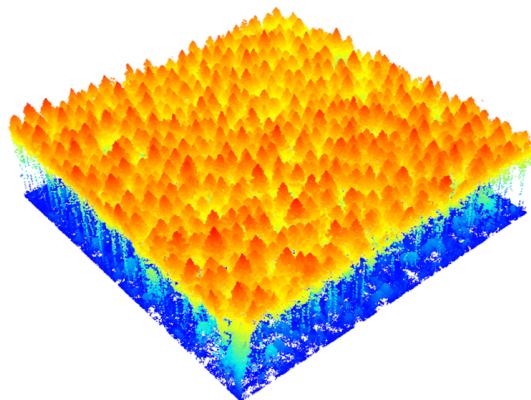


Figure 15.18: A normalized point cloud. (Du Toit, CC-BY-4.0).

## 15.26. CASE STUDY: CREATING LIDAR METRICS FROM A RAW POINT CLOUD387

The normalized point cloud is used to create our CHM. It is at this point that we can analyze the point cloud in a variety of ways. We can use an area-based approach (ABA) to create metrics at the grid level (using our normalized point cloud), or we can derive metrics at the individual tree scale. In order to do this we need to first segment the trees before creating metrics. Segmentation can be tedious, as it requires the user tweak the parameters of their segmentation algorithm. Understanding your forest type, and inspecting the point cloud visually can be very useful when deciding what kind of segmentation routine you want to run. Below we can see an example of a segmented point cloud.

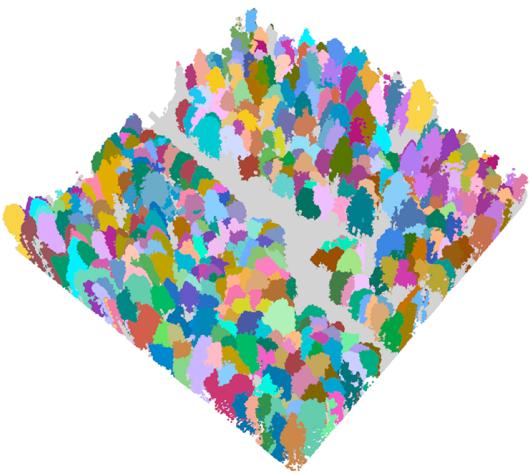


Figure 15.19: A segmented point cloud using the algorithm based on Dalponte and Coomes [2016]. (Du Toit, CC-BY-4.0).

Creating metrics at these two scales are the fundamental way in which we currently think about analyzing forests using LiDAR. Each tells us different things about the forest, so it is important to take scale into account when choosing what kind of analysis you want to do. Below we can see a Leaflet map of some area based metrics created for this point cloud!

### Your Turn!

In Figure 15.20 below, explore some of the LiDAR derivatives that we produced in the case study above. We can see how the pattern for maximum height follows the mean height quite closely. In addition, we can see that the 15th percentile of heights (ZQ 15) also broadly follows this pattern. However, we can see that the standard deviation of heights in the tall areas are quite variable, and that very little ground is visible in our tile (as the percentage of points above 2 m is very high almost everywhere).

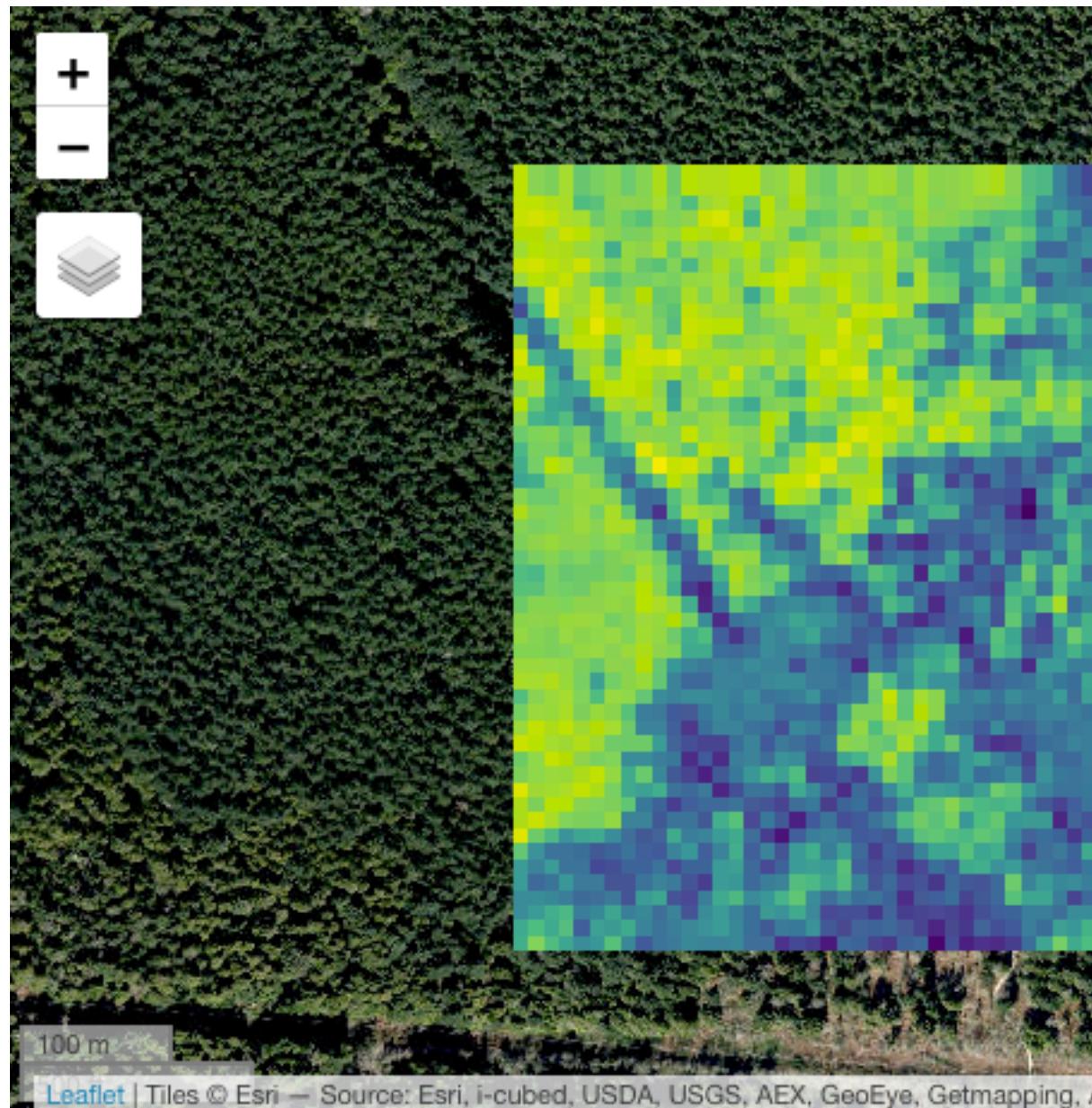


Figure 15.20: Explore different area-based metrics produced in the case study above. Hover over tile icon (top left) to turn different layers on or off. Du Toit, CC-BY-4. Animated figure can be viewed in the web browser version of the textbook: <https://www.opengeomatics.ca/LiDAR-acquisition-and-analysis.html#fig:5-LiDAR-derivatives-map>.

## 15.27 Summary

LiDAR is an active remote sensing technology that requires three components; a laser scanner, a GNSS, and an IMU. The laser scanner emits short, intense pulses of light and measures the time it takes for energy to be reflected back to the device. Time of flight principles are used to produce 3D point clouds, as we know exactly where in space the laser scanner is using the GNSS and IMU. These 3D point clouds can be used to monitor vegetation structure, stream properties, and topography. Point cloud derivatives such surface models can help us analyze different aspects of a forest or landscape. These surfaces can describe the ground (DEM), the top of the Earth's surface (DSM), or the height of the forest canopy (CHM) when the point cloud is normalized. When using LiDAR in a forestry context, it is important to decide on what resolution we need for our study; we can use an ABA or ITD approach to create LiDAR metrics, which can be used to make different types of inferences about the forest.

### Reflection Questions

1. What are the three main components of a LiDAR system?
2. Why is LiDAR so useful for forestry operations compared to other remote sensing technologies?
3. How is a DEM created? What are the main step that are needed?

### Practice Questions

1. If pulse of light is reflected off a surface 50 m away, what is the travel time of that pulse of light?
2. If a plane is flying at 450 m above the ground and the LiDAR sensor has a beam divergence of 0.17 mrad, what is the footprint diameter?
3. Explain what is meant by the term ‘discrete return’?
4. What is a typical LiDAR wavelength?

### Recommended Readings

- A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. Information Report FI-X-010 - White et al. [2013]
- Airborne laser scanning for quantifying criteria and indicators of sustainable forest management in Canada - Goodbody et al. [2021]
- Use of LIDAR in landslide investigations: A review - Jaboyedoff et al. [2012]



# Chapter 16

# Data Integration

Written by June Skeeter and Paul Pickell

**Data Integration,** This chapter is about some of the more practical aspects implementing GIS in a workflow. What type of problems and pitfalls might you encounter and how do you account for them?

This chapter will walk you through a number of things you will encounter when working in GIS, using an applied example as a guide. Then you will be presented with two other case studies showing you example workflow. One will show how the police involved deaths data presented in Chapter 3 was compiled. The second will show (forest stuff - second case study - need guidance on how to include) [Skeeter et al., 2022].

## Learning Objectives

1. Objective one
2. Objective two
3. Objective three

## Key Terms

Data, Integration, Other Stuff

### 16.1 Problems with Data Integration

Most GIS projects require us to analyze multiple data layers, sometimes from disparate sources to answer our research question. When working with different layers from different sources you are likely to encounter multiple incongruousness. What do you do if some of your layers are in vector format and some in raster?

What if one of your datasets is 10 years older than another? How do you handle data that were collected at different resolutions or scales stored in different file types? These are questions that pop up every day when working in GIS.

We will discuss what to do when you encounter different: 1) Data types, sources, formats  
2) Data resolutions 3) Datum, extents, scales 4) Time periods, collection dates

## 16.2 Framing The Problem

For millennia, wetlands in the Canadian Arctic have been accumulating large stockpiles of Carbon. Permafrost (frozen ground) and short growing seasons in these landscapes cause dead organic matter freezes into the soil before it can fully decompose. Climate change in the Arctic is causing permafrost to degrade, at rapid rates in some regions. This will speed up decomposition of these large Carbon stockpiles and could potential result in a large pulse of greenhouse gasses (Carbon Dioxide and Methane) being released into the atmosphere. Creating a positive feedback mechanism that further exacerbates warming. At the same time, climate change is causing trees and shrubs to encroach on the tundra and leading to longer growing seasons. Increased plant growth sequesters Carbon Dioxide and serves as a negative feedback mechanism (“Remediating” climate change).

[Find/make and insert diagram]

Monitoring Carbon balances of Arctic ecosystems is especially difficult and expensive due to the harsh conditions and inaccessible nature of most locations. Because of this, very little is known about how these systems are and will continue responding to climate change relative to other biomes. Figuring out the Carbon balance of these can help fill a “big knowledge gap” and improve the accuracy of global climate models.

## 16.3 About The Data

The MacKenzie Delta (xx km<sup>2</sup>) is the second largest Arctic Delta in the world. It is a patchwork of river channels, lakes, boreal forests, and Carbon rich wetlands. It is very much understudied from the perspective of climate sciences. To date, only one ground based observation of landscape level Carbon exchange has been made anywhere in the Mackenzie Delta. In 2017, an intensive study was conducted using a method known as Eddy Covariance to measure the uptake/emission of Carbon Dioxide and Methane across a wetland (xx m<sup>2</sup>) in real-time (30-minute intervals) at a site in the Delta known as Fish Island [Skeeter et al., 2022].

In the site was a strong sink for Carbon during the growing season in 2017. But, Arctic climates are characterized by extreme inter-annual variation so one year



Figure 16.1: The Eddy Covariance system at Fish Island [Skeeter et al., 2022].

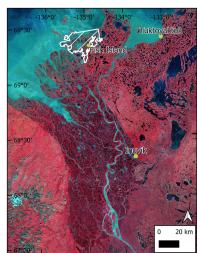


Figure 16.2: Maybe insert Web-map showing site instead. [Skeeter et al., 2022]

alone cannot be used determine the average carbon balance. Unfortunately, due to funding issues, what was supposed to be a multi-year research campaign, was shutdown after just one season.

Regardless ...

How can we use this one year of data from one point location to get a better idea of the Carbon balances in the Arctic? We can pull in data from other sources, do a bit of fancy modelling, and a few “back of the envelope” calculations to come up with some ballpark guesses.

## 16.4 Data Resolution

What do you do if your data are collected at different resolutions?

## 16.5 Integrating Vector and Raster Data

How can you work with both raster and vector data and when might you want to switch between data types?

Evey Eddy Covariance observation has a “footprint” or upwind source area for the Carbon. It is calculated using some complicated calculus that is well beyond the scope of this class, but

## 16.6 Rasterization

Say you have a vector layer of landscape classification scheme and need to intersect it with a source area raster

## 16.7 Vectorization

Say you have a model that outputs a raster layer representing an upwind source area for an Eddy Covariance observation and you want to display it in a more human friendly format.

## 16.8 Zonal Statistics

Say you have a raster layer (e.g. maximum annual NDVI) and you want to describe it over a certain region.

## 16.9 Smoothing

### 16.10 Simplifying

### 16.11 Spatial Data Errors

### 16.12 Accuracy vs. Precision

Measurement Errors Accuracy: The degree to which a set of measurements correctly matches the real world values. How close are we to the real value? If there is a consistent (systematic) offset from that real world value, our measurements are inaccurate. They have a bias. Precision: The degree of agreement between multiple measurements of the same real world phenomena. How repeatable is a measurement? If you take five measurements of the same feature, how likely are they to be similar? Lack of precision can be attributed to random errors.

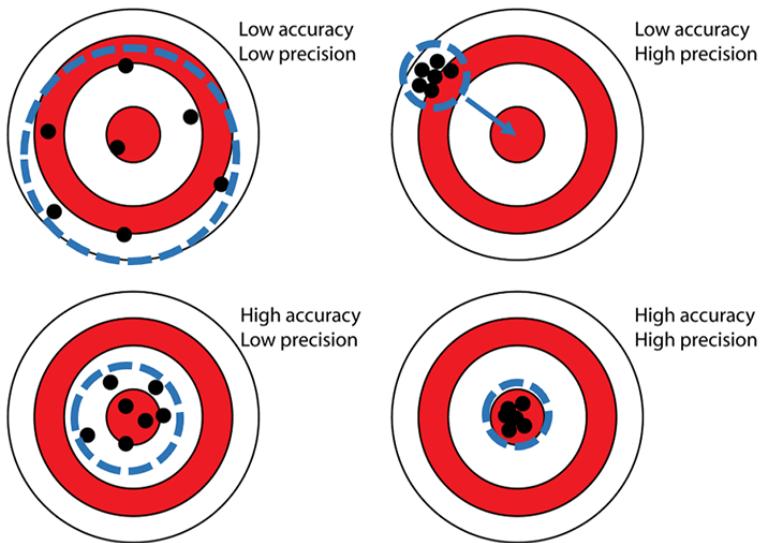


Figure 16.3: Precision versus accuracy [Davies, 2020b].

### 16.13 Vagueness and Ambiguity

Vagueness - Victoria ... does it mean Victoria BC vs. Victoria AU

Ambiguity - coastline - is it the high water line? Low water line? mean water level?

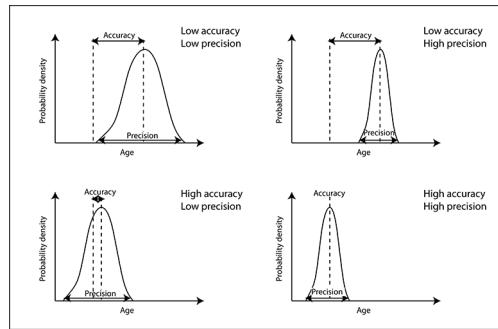


Figure 16.4: Accuracy and precision in chronostratigraphy [Davies, 2020a].

## 16.14 Quantifying Spatial Errors RMSE, Euclid's Distance

## 16.15 Logical Errors

Data incongruousness

## 16.16 Ecological Fallacy, Atomistic Fallacy, MAUP etc. Its important to include these, whether here or elsewhere?

## 16.17 Other Errors?

- source data errors, out of date data, data entry & digitization?

## 16.18 Case Study: Title of Case Study here

You see textual case study content here

### 16.18.1 Large Scale

Footprint mapping, temporal upscaling. I'll fill in more text here later, these figs are just grabbed from my thesis chapters. The gist of it - Measured NEE in one year. Have 10 years of climate data + Reanalysis data + satellite data. Combine these data sources & train a model to do a temporal upscale/sensitivity analysis to see how inter-annual climate variability impacts NEE. Then do a landscape classification with a greenest pixel NDVI image, intersecting with the flux footprint. Use that to find the representative areas to do a “back of the envelope” spatial upscaling.

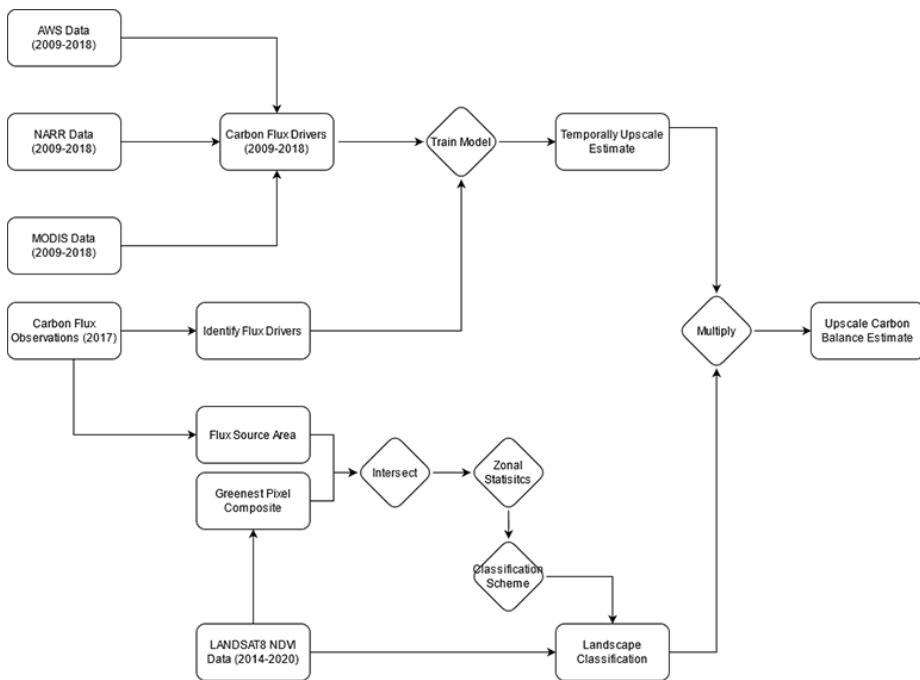


Figure 16.5: Rough flowchart draft.

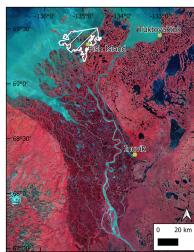


Figure 16.6: Reference map showing the Mackenzie Delta (Currently from chapter 2, I'll change it to a full delta NDVI map).

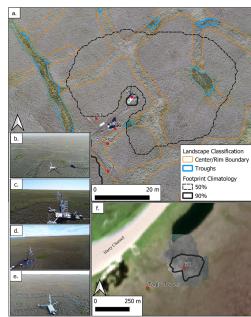


Figure 16.7: Landscape classification and drone imagery.

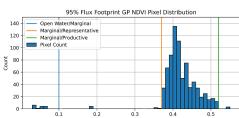


Figure 16.8: Footprint NDVI profile.

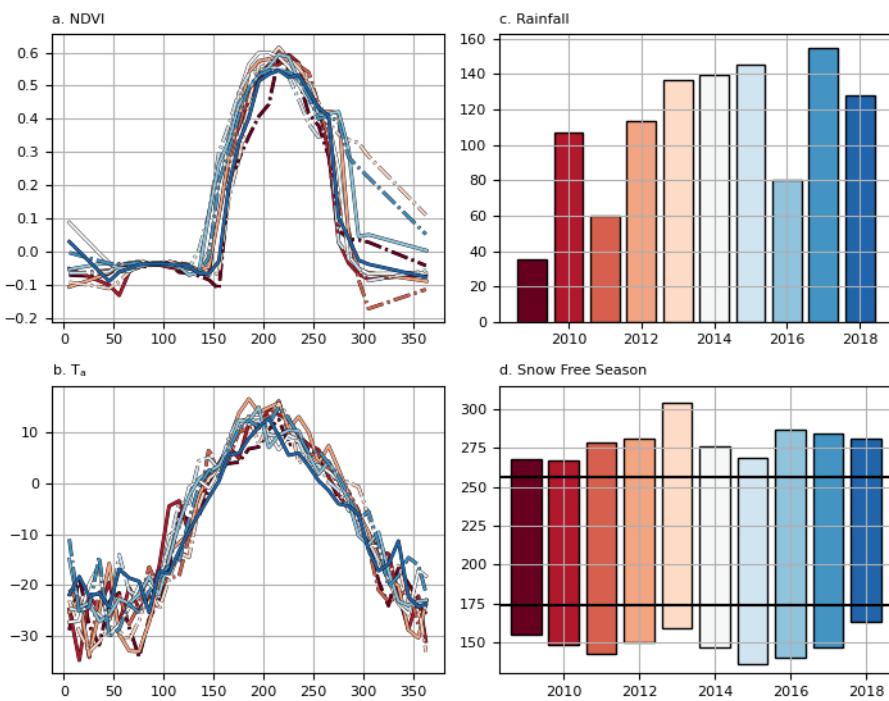


Figure 16.9: Climate Data.

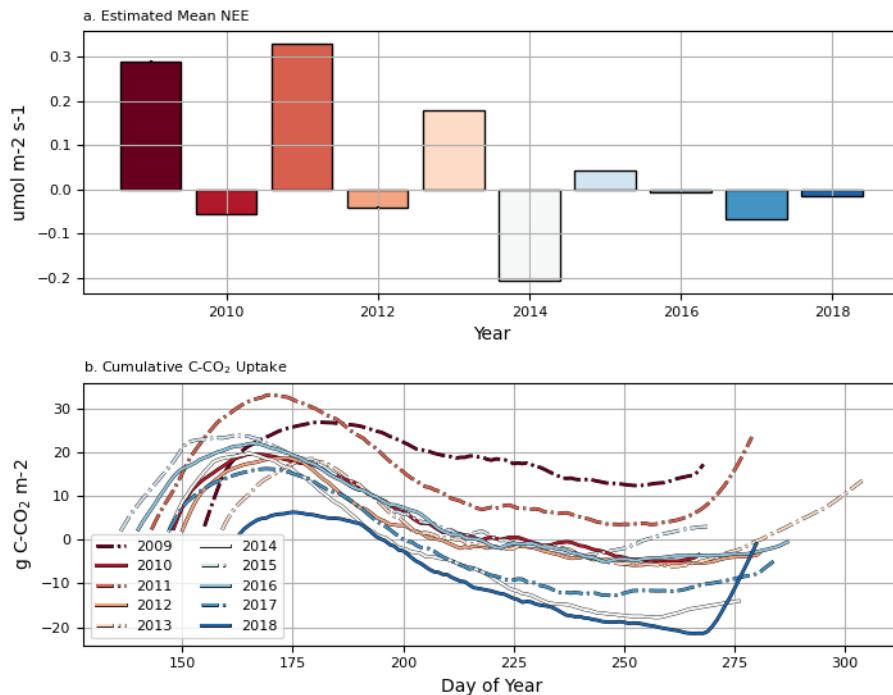


Figure 16.10: Temporally upscaled flux estimate.



Figure 16.11: Landscape classification.



# Chapter 17

## Making Beautiful Maps

Written by Maya Daurio and Paul Pickell

A map *is not* the territory it represents, but, if correct, it has a *similar structure* to the territory, which accounts for its usefulness. — Alfred Korzybski [1933]

**Making beautiful maps**, or *making effective maps*, is comparable to being able to tell a good story. Just like a good story, a beautiful map comprises certain elements that, put together, frame the narrative in a way that makes it easy for others to understand and interpret. In this way, a map is a communication device. It conveys information through a visual representation of the spatial relationships of the features for a specific area. To better understand maps as particular form of visual communication, it may be helpful to think about what it is that maps actually do.

Maps represent areas larger than we can see, usually from above, and some depict phenomena that cannot be seen [Manson and Matson, 2017], such as growing seasons. They also illustrate spatial relationships, such as this map showing the percentage of the population aged 65 and over by census tract in Moncton, New Brunswick. We can deduce that those aged 65 years and over are concentrated in certain areas of the city, using census tracts as a proxy.

**Cartography** is the art and science of designing maps and consists of certain principles and rules that help ensure accuracy and clarity. All maps are simplified representations of a place and reflect the particular choices and motivations of the cartographer. Cartography is as much about deciding what to omit from a map as it choosing what to include. Making beautiful maps involves the art, and science, of selecting and modifying data and portraying concepts with clarity and precision. Decisions around colour, font size, graphical hierarchies, classification themes, and legends are all elements of maps that determine how effective, and accurate, they are in conveying information. In this chapter, you will learn about what elements go into making an effective map, the different

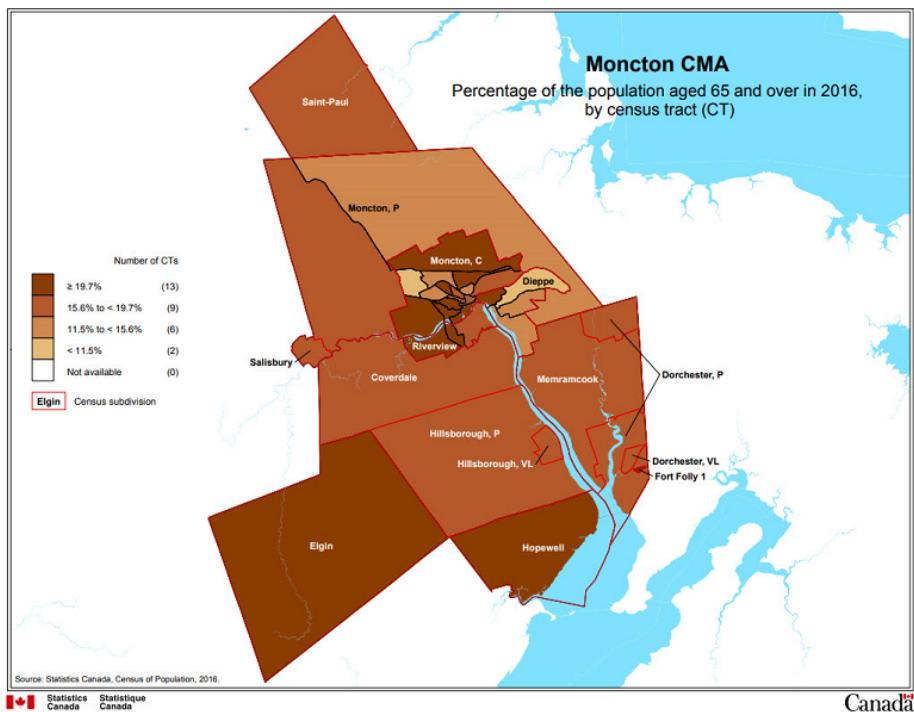


Figure 17.1: Maps illustrate how variables are related in space. [Statistics Canada, 2016] (c) Statistics Canada, Open Government Licence - Canada.

categories of maps, and the principles of map design.

## Learning Objectives

1. Identify elements of cartographic design
2. Describe different types of maps
3. Describe different types of symbology and classification and their uses

## Key Terms

Map elements, symbology, classification, cartography, north arrow, scale, thematic, choropleth, dot density, reference, isoline

### 17.1 Types of Maps

There are many different types of maps. Just try doing a Google search, and you will come across descriptions of thematic, cadastral, topographic, and physiographic maps, to name but a few. There is no standard agreement on how many different categories of maps exist. Although some differentiate among as many as five types of maps based on the functions they serve [ICSM, 2021], all maps can also be categorized into just two types, **reference** and **thematic**. Reference maps represent the human-made or natural features of the landscape and are sometimes thought of as basemaps. They provide information about a particular location, such as in the example below, depicting the longest rivers in the location of Canada and where they overlap with the United States. Topographic maps based on the National Topographic System (NTS) are another example of reference maps.

### 17.2 Thematic Maps

Thematic maps depict the spatial distribution of particular features, which are symbolized according to the quantitative or qualitative values of their attributes. Thematic maps, as the name implies, emphasize a particular theme, using qualitative or quantitative data. Whereas reference maps show the locations of things, such as the locations of mountains, a thematic map might represent the geology of those mountains.

Qualitative thematic maps illustrate the spatial extent of categorical data [Anderson, 2020b], such as the 1956 map bedrock geology map below.

Quantitative thematic maps depict the spatial patterns of numerical data, often expressed as rates or percentages, as in the map below showing the difference in population change between two periods, where the difference is expressed as a percentage.

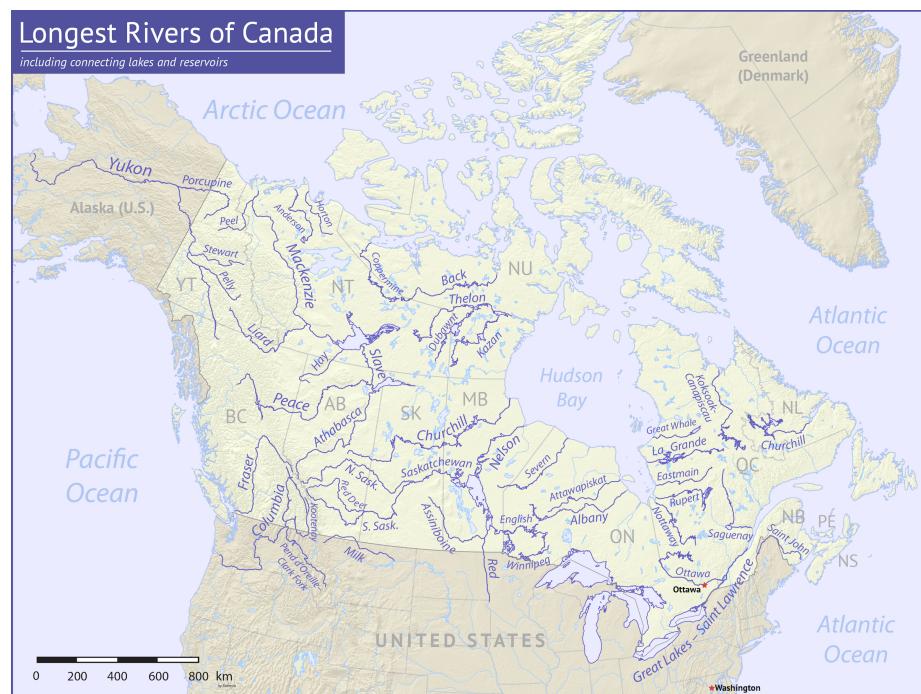


Figure 17.2: One example of a reference map. [Shannon1, 2017] (c) Shannon1, CC BY-SA.

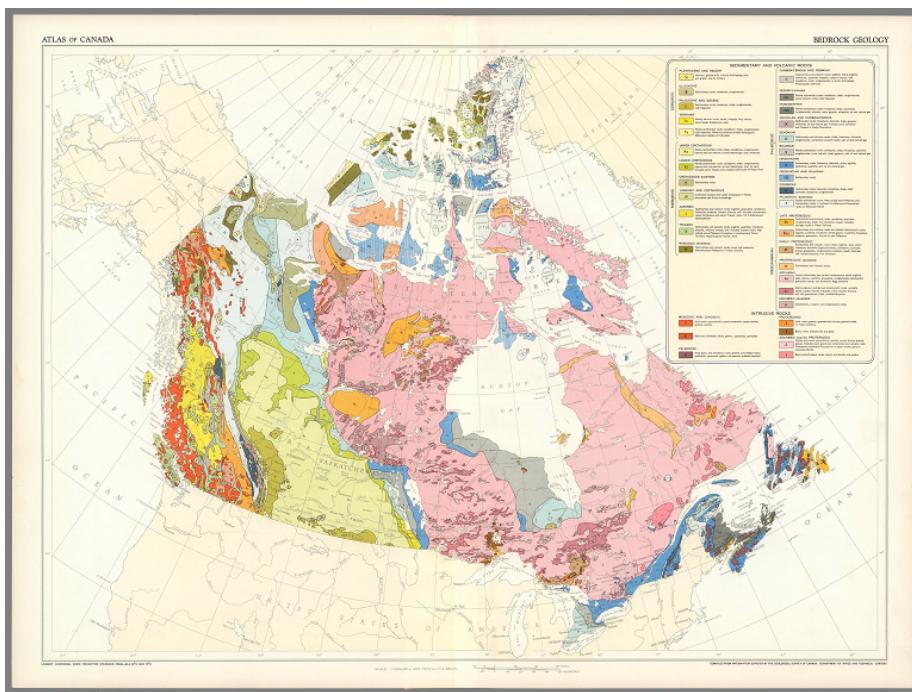


Figure 17.3: A thematic map using qualitative categorical data. [Nicholson and Comtois, 1956] (c) Comtois and Nicholson, CC BY-NC-SA 3.0.

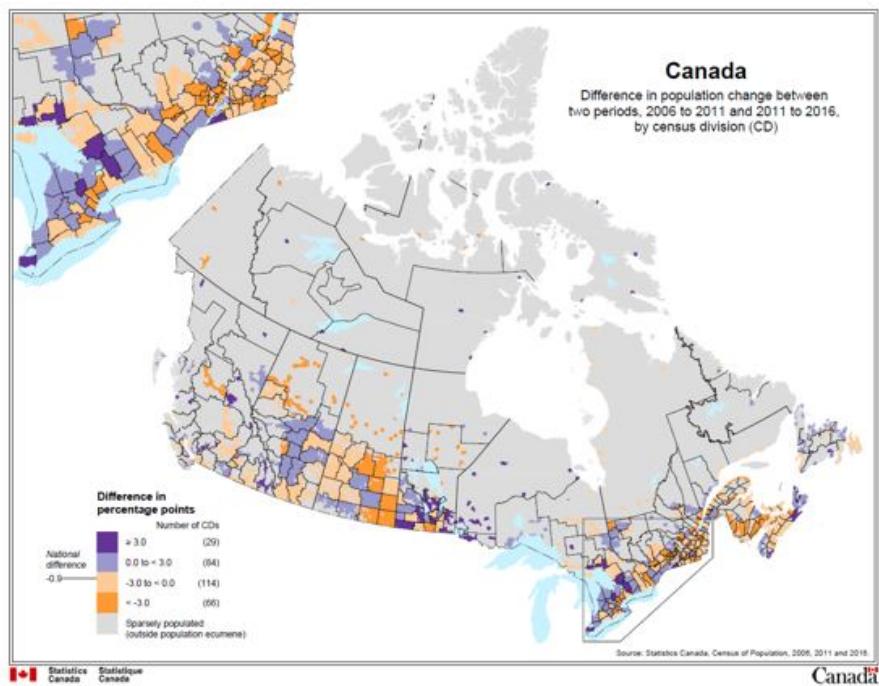


Figure 17.4: A thematic map depicting population data as a percentage. [Statistics Canada, 2017] (c) Statistics Canada, Open Government License - Canada.

## 17.3 Choropleth Maps

There are many different types of thematic types. A commonly used thematic map is a **choropleth** map, which applies different colours or shading to the entire extent of a predefined areal unit [Weiss et al., 2008]. Data used in choropleth maps must be standardized and represented as rates or ratios rather than raw counts or numbers.

One of the reasons choropleth maps are so widely used is because much of the data with which we're concerned, such as demographic data, is tied to areal units like census tracts or counties.

The map below represents, using percentages instead of absolute numbers, the spatial distribution of the percentage of Cantonese speakers per total population in each census tract in the city of Vancouver, British Columbia. Household population of mother tongue language speakers by census tract is an example of demographic data collected by Statistics Canada and can serve as a proxy for linguistic diversity. This map shows greater concentration of Cantonese speakers in census tracts in the East Vancouver area.

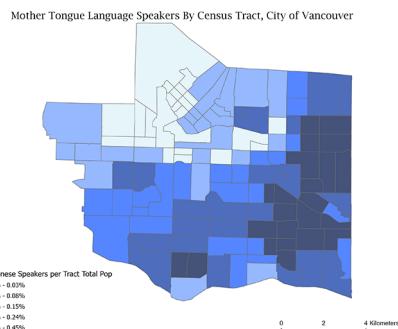


Figure 17.5: A choropleth map representing the percentage of Cantonese speakers per total population of census tracts in the City of Vancouver Household Population by Mother Tongue retrieved from the web application SimplyAnalytics.

This map is also a good example of the potential problems with choropleth maps, which has to do with something called the Modifiable areal unit problem, or MAUP. This results when data such as population, which is the measurement of population per unit area, is aggregated into an areal unit, where the shape and scale of the unit affects the resulting rates, or values, of the data. In this map example, the areal unit is a census tract, which vary by size and population density across the city and are delineated according to factors other than number of mother tongue language speakers. Without knowing anything about census tracts, we may assume based on this map that an entire census tract contains a

disproportionate percentage of Cantoneses speakers, when in reality it may be just a few households more than the census tract next to it. This map fails to convey that kind of nuance and illustrates one of the weaknesses of choropleth maps, in spite of their wide use and applicability, particularly when displaying census data.

## 17.4 Dot Density Maps

In contrast to choropleth maps, raw data/counts (e.g. number of mother tongue language speakers) as well as rates/ratios (e.g. number of mother tongue language speakers per square kilometer) can be used in **dot density** maps. These maps use a one-to-one density, where one dot represents one count, or a one-to-many density, where one dot represents a number of counts to map the spatial distribution of a certain phenomenon. In the map below, one dot represents 20 speakers, and rather than mapping one language community, we can view the spatial distribution of two language communities to better understand different settlement patterns between two groups of Chinese speakers.

Also unlike choropleth maps, dot density maps do not need to be tied to enumeration units, although they can be, such as in this map example, which uses another method of mapping the same data that was used in the choropleth map above. For their part, dots are distributed randomly, and a potential misleading aspect of dot density maps is the false interpretation that the dots are placed in precise locations in space.

## 17.5 Isoline Maps

- 17-isoline.png from row 23 of OER Content Sharing

## 17.6 Diagrammatic Maps

*Note: I think I would combine diagrammatic and cartogram subsections.*

## 17.7 Cartograms

- 17-cartogram.png from row 42 of OER Content Sharing

*Note: Paul, see example interactive quiz at bottom of this Introduction to Geomatics OER textbook.*

## 17.8 Additional Resources on Types of Maps

- Thematic Maps
- Choropleth Maps

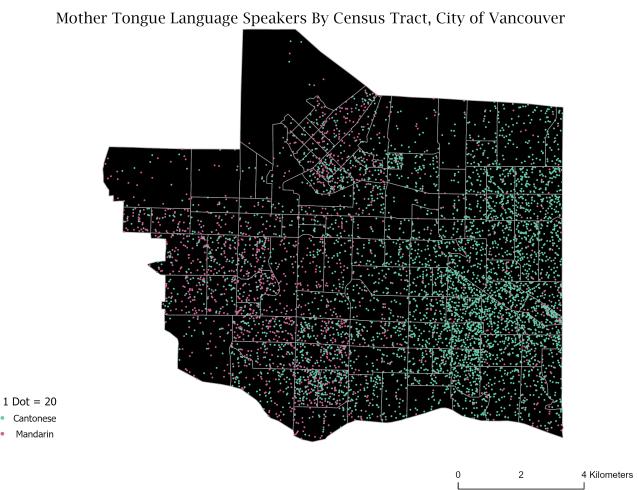


Figure 17.6: Dot density map showing the distribution of Cantonese and Mandarin mother tongue language speakers by census tract Household Population by Mother Tongue retrieved from SimplyAnalytics.

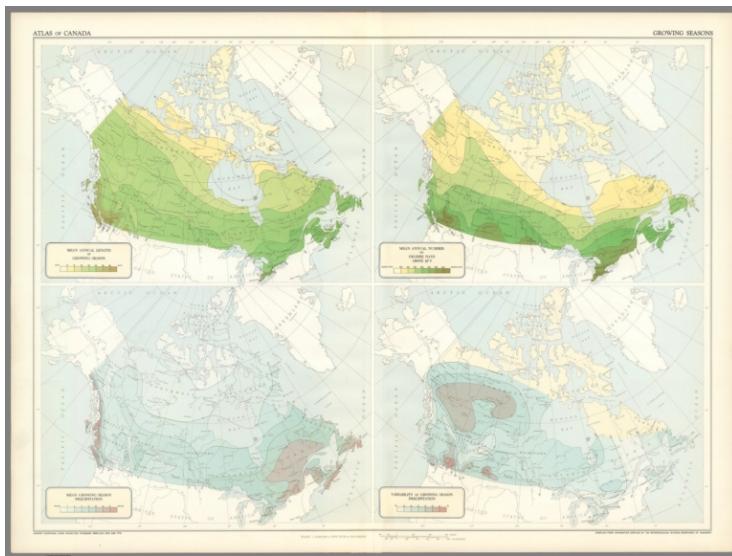


Figure 17.7: An isoline map representing growing seasons from the *Atlas of Canada*. Isolines are lines which connect points with identical values. ('Growing Seasons', [Nicholson and Comtois, 1950], CC BY-NC-SA 3.1) retrieved from David Rumsey Map Collection.

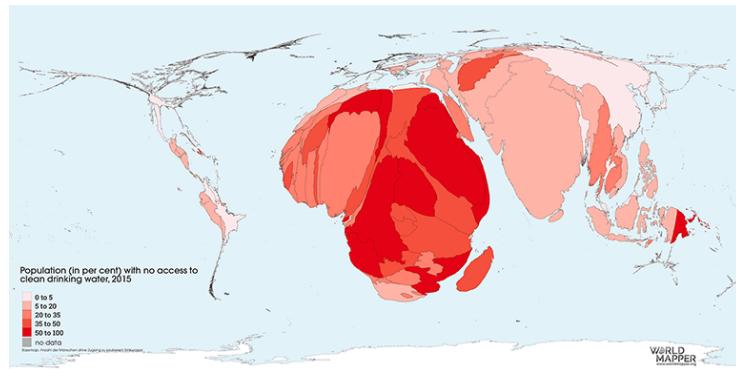


Figure 17.8: An effective illustration of a cartogram representing the population (%) with no access to clean drinking water. ([WorldMapper, 2015], CC BY-NC-SA 4.0).

- Dot Density Maps
- Dot density map
- Types of Maps

## 17.9 Map Composition

Most effective maps comprise standard elements that help orient the map viewer in space and provide helpful contextual information. There is no one way to compose a map, but a good cartographer will arrange and design map elements to provide the greatest clarity. Many maps not only include the phenomenon or object being mapped, called the **figure**, but also the spatial context of the figure, representing where it is in relation to the space around it, called the **ground**. Maps often have a **frame**, which bounds the figure and any other map elements within a predetermined length and width. Some frames are visible to the map reader, while others are not and simply act as a bounding box for the cartographer while designing the map.

## 17.10 Figure

- 17-figure.png from row 32 of OER Content Sharing

## 17.11 Ground

- use 17-figure.png

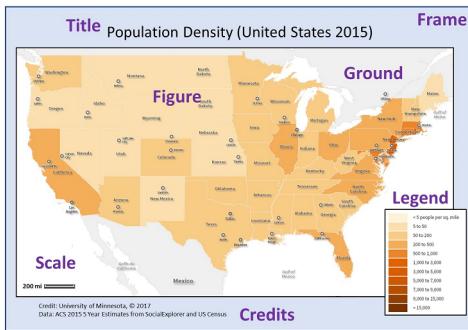


Figure 17.9: The components of effective map composition. ('Map composition.', Manson [2017], CC BY-NC-SA 4.0)

## 17.12 Frame

- use 17-figure.png

## 17.13 Elements of Maps

Whereas the figure is the most prominent element of a map, most maps also contain additional elements that provide important supplementary information, without which the purpose of the map would be less clear. Thinking back to the dot density example from the previous section, it would be difficult to understand what the map was about without a title and a **legend**, which contained labeling and symbols explaining the units in the map.

Other elements are important for spatial orientation, such as the **scale** and **north arrow**, which provide distance and directional information, particularly critical for navigational maps such as a park trail map.

Just as in writing, it is also important to provide information about the data sources used in creating the map. Not only does this credit the authors of the data, it also provides transparency and the ability to assess the reliability of the data.

## 17.14 Text

- 17-text.png from row 43 of OER Content Sharing

## 17.15 Legend

- can use 17-figure.png

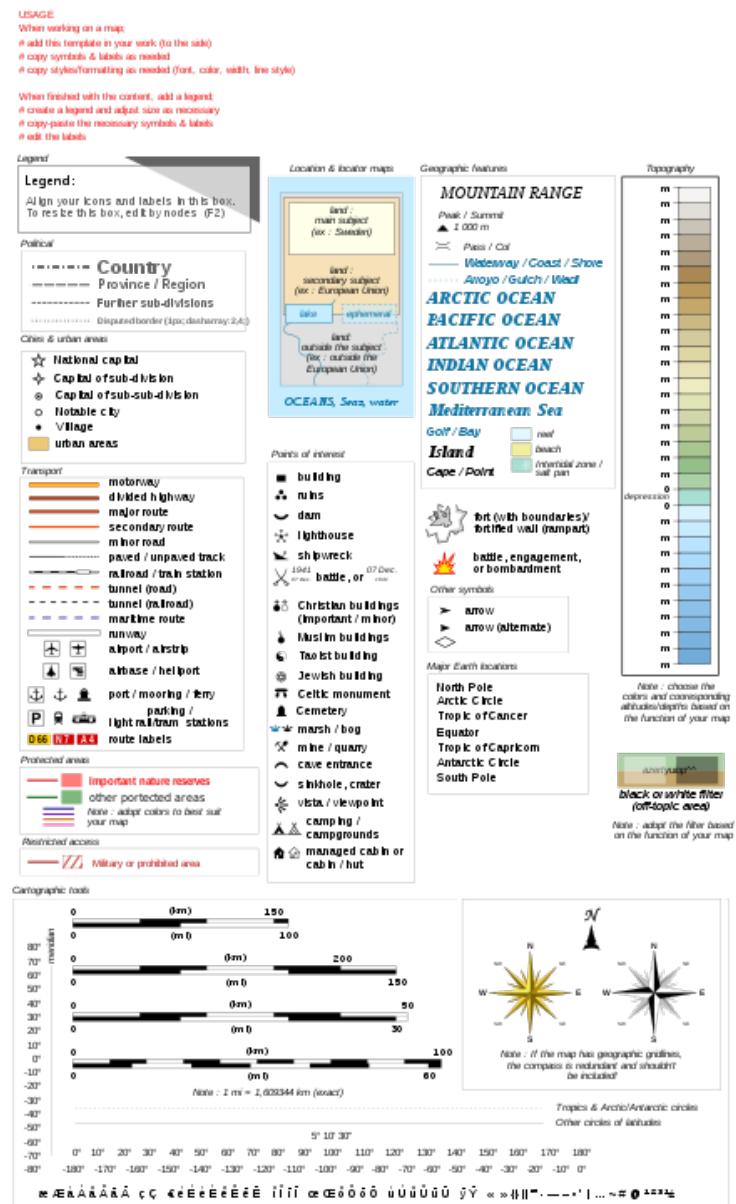


Figure 17.10: Typography standards for cartography. ('Cartographic style guide', Gaba [2014], CC BY 2.5).

## 17.16 Scale and North Arrow

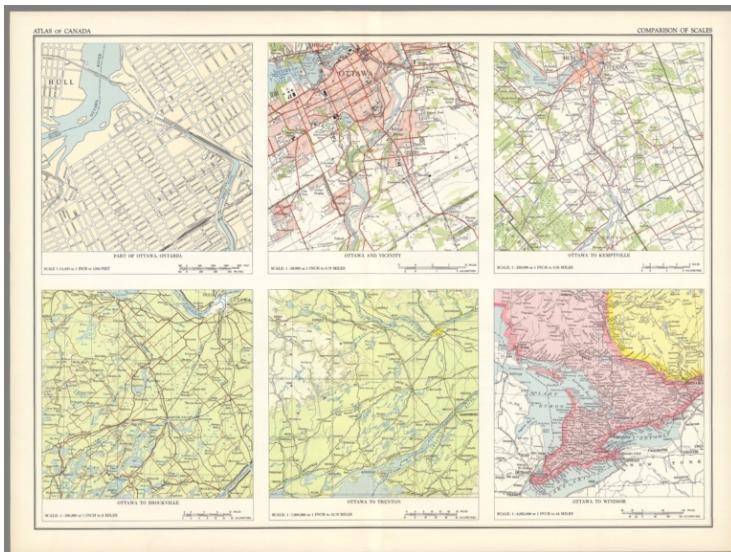


Figure 17.11: Six maps of Ottawa at different scales from the *Atlas of Canada*. ('Comparison of scales', Nicholson and Comtois [1958], CC BY-NC-SA 3.1).

- 17-scale.png from row 21 of OER Content Sharing
- image is great example of both concept of scale and scale bars, not of north arrow

## 17.17 Measured Grid

- see Wikipedia entry referenced in row 41 for two possible images

## 17.18 Citation

- – can use 17-figure.png

## 17.19 Symbolization

**Symbolization** is the process by which cartographers choose how to represent the features on a map. Unless the map comprises an aerial or satellite image of the surface of the earth, all maps are representations of reality, whereby information is conveyed through the use of symbols.

There are many ways to visually alter symbols - represented by points, lines, or polygons - to convey information and show both qualitative and quantitative

differences between features in the map. These include varying symbols by size, shape, and colour, to name just a few. Using points to map the location of Canada's cities, for example, the cartographer may use different shapes to illustrate differences between cities and towns. To show variations in populations of those cities, the cartographer might vary the size of the symbols.

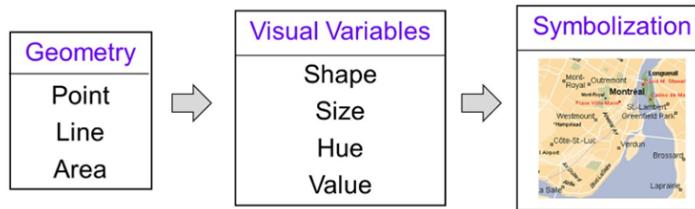


Figure 17.12: The components of symbolization. ('Symbolization', Manson [2013b], CC BY-NC-SA 4.0).

## 17.20 Separable

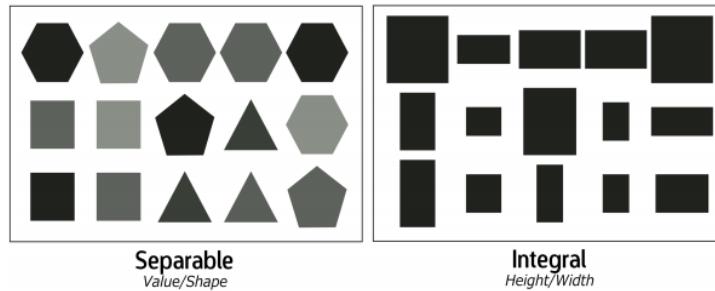


Figure 17.13: Separable and integral visual combinations delineated by value/shape and height/width, respectively. (Elmer [2012])

- 17-separable.png from row 33, page 7 of OER Content Sharing

## 17.21 Integral

- use 17-separable.png for this too

## 17.22 Graduated

- 17-graduated.png from row 45 of OER Content Sharing

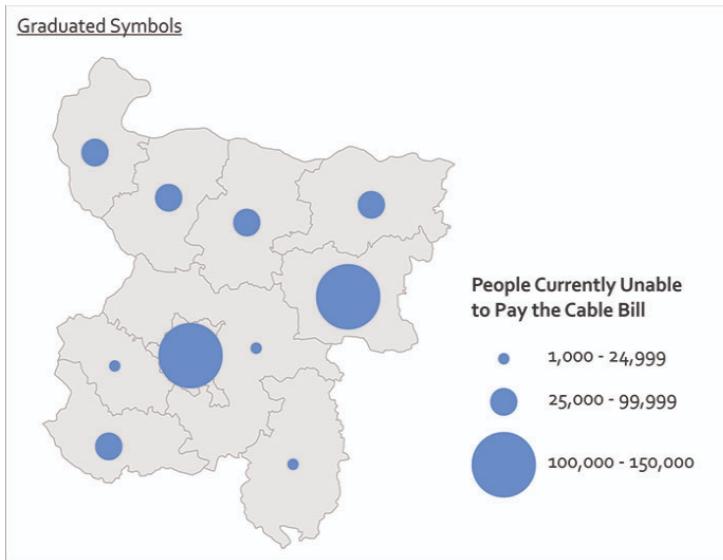


Figure 17.14: Graduated symbols vary symbol size representing a category of values. ('Proportional & Graduated Symbols', Robinson [2020], CC BY-NC-SA 4.1).

## 17.23 Configurable

- 17-configurable.png from row 33 of OER Content Sharing

## 17.24 Proportional

- 17-proportional.png from row 45 of OER Content Sharing

## 17.25 Line Weight

## 17.26 Additional Resources

- The Graduated Colour Map: A Minefield for Armchair Cartographers

## 17.27 Colour

Cartographers use **colour** in maps to communicate information, to differentiate among features, and to illustrate spatial patterns. Maps are interpreted visually, so cartographic decisions about colour are particularly important.

Colour can also influence the feeling of a map. Some cartographers caution

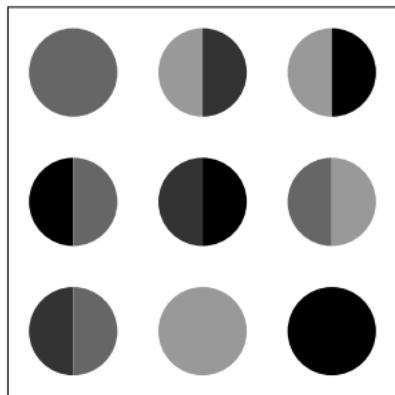


Figure 17.15: Separable and integral visual combinations delineated by value/shape and height/width, respectively. (Elmer [2012])

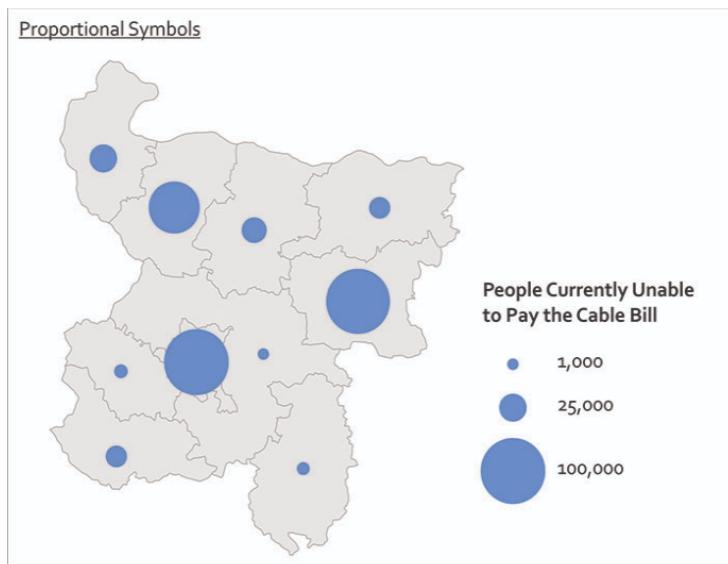


Figure 17.16: Proportional symbols vary size of symbols in relation to their attribute value. ('Proportional & Graduated Symbols', Robinson [2020], CC BY-NC-SA 4.1.)



Figure 17.17: Varying the thickness of different lines in the map can help focus attention and differentiate between map features. ('Map Elements and Design Principles', Ingram [2021], CC-BY-SA-4.0).

against using red for symbols, for example, because it can convey a sense of alarm. In general, it is a good idea to pay attention to how colours convey intensity.

Choice of colours should also take into account that some people are colourblind, especially when choosing a colour scheme. The online tool Colorbrewer 2.0 helps cartographers evaluate how easy it is to differentiate among colours in a colour scheme.

Colour is actually comprised of three elements: hue, chroma, and lightness, or value.

## 17.28 Hue

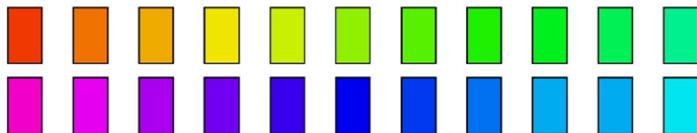


Figure 17.18: Hues are what we commonly think of as the names of colours and distinguish between qualitative data. ('Hue', Manson [2013a], CC BY-NC-SA 3.0).

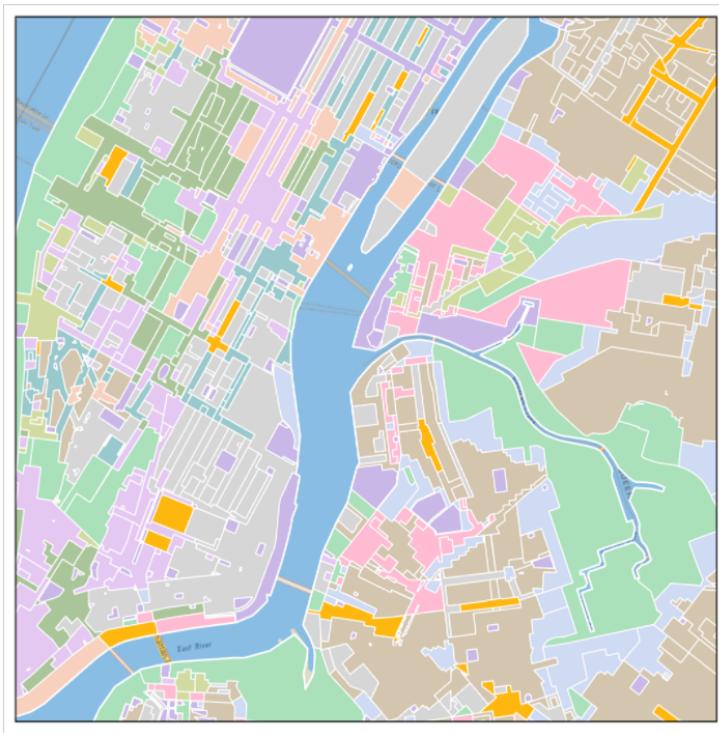


Figure 17.19: Higher values represent lighter colors while lower values refer to dark colors. ('Chroma', Anderson [2020a], CC BY-NC-SA 4.0).

## 17.29 Chroma

### 17.30 Lightness

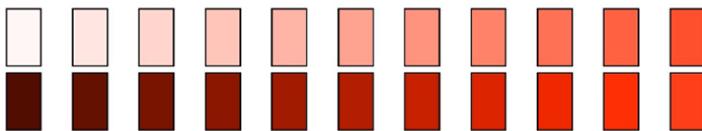


Figure 17.20: Higher values represent lighter colours while lower values refer to darker colours. ('Value', Manson [2013c], CC BY-NC-SA 4.0).

### 17.31 Bivariate Colour Schemes

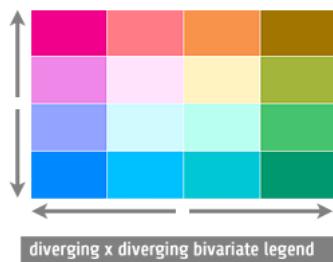
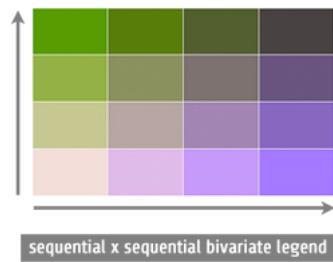


Figure 17.21: Different examples of bivariate legends. ('Bivariate', AxisMaps [2020], CC BY-NC-SA 4.0).

### 17.32 Colour Pickers

- maybe just reference ColorBrewer?

### 17.33 Additional Resources

- Colour Theory and Cartography
- Mapping COVID-19: How maps make us feel
- ColorBrewer 2.0

### 17.34 Classification Schemes

Data **classification** is the process of aggregating large numbers of observations into categories of values or data ranges. Classification is typically used for choropleth or graduated symbol maps. The purpose of classifying data is to simplify it for easier interpretation by the map reader. Rather than differentiating each value by modifying the symbol hue or size, values are grouped together into discrete classes.

There are many classification schemes which determine how breaks in the data are defined, and it can sometimes be difficult to understand which classification method is best to use or whether to use classification for your data at all. Choosing the best classification scheme depends on your data, what you are trying to communicate, and who is your intended map audience.

#### 17.35 Qualitative

#### 17.36 Sequential

#### 17.37 Intervals

#### 17.38 Quantiles

#### 17.39 Natural Breaks (Jenna)

#### 17.40 Standard Deviation

#### 17.41 Additional Resources

- Better Breaks Define Your Map’s Purpose
- Simplification

## 17.42 Generalization

One of the cartographer's biggest challenges in making a map is deciding what information to exclude. Because maps are a representation of reality that must fit within certain dimensions and communicate clearly to the map audience, the amount of detailed information included must be limited. This process of simplifying detailed information is known as **generalization**.

Cartographers have developed many techniques for eliminating, highlighting, or subduing visual information in maps, some of which have been incorporated into geoprocessing tools commonly used in GIS software. For example, merging two or more polygons together can be performed by running a tool.

### 17.43 Select

### 17.44 Amalgamate

### 17.45 Exaggerate

### 17.46 Displace

### 17.47 Refine

### 17.48 Simplify

### 17.49 Aggregate

### 17.50 Typify

### 17.51 Smooth

### 17.52 Enhance

### 17.53 Collapse

### 17.54 Merge

## 17.55 Additional Resources

-Cartographic generalization

## 17.56 Map Design

**Map design** is the process of incorporating cartographic design principles to produce an interesting and effective map. It involves arranging appropriate map elements, such as a title and legend, in a way that is easy for the map audience to interpret, as well as visually foregrounding more important information so that it stands out against less relevant details. Of course good map design is also about making effective choices for symbol colours and situating the map in such a way that the map audience understands its orientation in space.

Good map design effectively integrates map elements with design principles and symbolization to produce a persuasive map [Deluca and Bonsal, 2017].

## 17.57 Subject

## 17.58 Projection and Orientation

## 17.59 Hierarchy

- hierarchy.png from row 44 of OER Content Sharing for labeling hierarchy
- see Visual Hierarchy for discussion of and examples of visual hierarchy instead

## 17.60 Balance

## 17.61 Summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut in dolor nibh. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent et augue scelerisque, consectetur lorem eu, auctor lacus. Fusce metus leo, aliquet at velit eu, aliquam vehicula lacus. Donec libero mauris, pharetra sed tristique eu, gravida ac ex. Phasellus quis lectus lacus. Vivamus gravida eu nibh ac malesuada. Integer in libero pellentesque, tincidunt urna sed, feugiat risus. Sed at viverra magna. Sed sed neque sed purus malesuada auctor quis quis massa.

## Reflection Questions

1. Explain ipsum lorem.
2. Define ipsum lorem.
3. What is the role of ipsum lorem?
4. How does ipsum lorem work?

## Practice Questions

2. Given ipsum, solve for lorem.
3. Draw ipsum lorem.

## Recommended Readings

Anderson, C. (2020). Types of Maps. Retrieved June 11, 2021, from *GEOG 486: Cartography and Visualization* website: <https://www.e-education.psu.edu/geog486/node/641>

Cote, Paul. (2021). GIS Manual: Elements of Cartographic Style. *PbcGIS*. <https://www.pbcgis.com/style/>.

Deluca, E., & Bonsal, D. (2017). Design and Symbolization. In S. Manson (Ed.), *Mapping, Society, and Technology*. Retrieved from <https://open.lib.umn.edu/mapping/chapter/4-design-and-symbolization/>

Intergovernmental Committee on Surveying and Mapping (ICSM). 2021. "Types of Maps." *Overview to the Fundamentals of Mapping*. <https://www.icsm.gov.au/education/fundamentals-mapping/types-maps>.

Korzybski, A. (1933). *Science and sanity: an introduction to non-Aristotelian systems and general semantics* (1st ed.). Retrieved from [https://openlibrary.org/books/OL24876034M/Science\\_and\\_sanity](https://openlibrary.org/books/OL24876034M/Science_and_sanity)

Manson, S., & Matson, L. (2017). Maps, Society, and Technology. In S. Manson (Ed.), *Mapping, Society, and Technology*. Retrieved from <https://open.lib.umn.edu/mapping/chapter/1-maps-society-and-technology/>

Thomas, I. (2001). Thematic cartography today: recalls and perspectives. *Cybergeo: European Journal of Geography*, 189. <https://doi.org/10.4000/cybergeo.34958>

Weiss, C., Cillis, P., & Rothwell, N. (2009). Population Ecumene of Canada: Exploring the Past and Present. *Geography Working Paper Series*, 2008(003).



# Bibliography

- Amir Ajourlou. Introduction to Network Models, 2018. URL <https://ocw.mit.edu/courses/civil-and-environmental-engineering/1-022-introduction-to-network-models-fall-2018/>. Publication Title: MIT OpenCourseWare Type: Course.
- Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *In Proceedings of the 2nd International Symposium on Information Theory*, 1973. doi: 10.1007/978-1-4612-1694-0\_15.
- American Society for Photogrammetry & Remote Sensing. LASER (LAS) FILE FORMAT EXCHANGE ACTIVITIES, 2019. URL <https://www.asprs.org/divisions-committees/lidar-division/laser-las-file-format-exchange-activities>.
- Cary Anderson. Chroma, 2020a. URL <https://www.e-education.psu.edu/geog486/node/606>.
- Cary Anderson. Types of Maps. In *GEOG 486: Cartography and Visualization*. The Pennsylvania State University, 2020b. URL <https://www.e-education.psu.edu/geog486/node/641>.
- Luc Anselin and Anil Bera. Spatial Dependence in linear Regression Models with an Introduction to Spatial Econometrics. In A Ullah and D Giles, editors, *Handbook of Applied Economic Statistics*, pages 237–289. Marcel Dekker, New York, 1998. doi: 10.1201/9781482269901-36. Section: 7.
- Eric Asa, Mohamed Saafi, Joseph Membah, and Arun Billa. Comparison of Linear and Nonlinear Kriging Methods for Characterization and Interpolation of Soil Data. *Journal of Computing in Civil Engineering*, 26(1):11–18, 2012. ISSN 0887-3801. doi: 10.1061/(asce)cp.1943-5487.0000118.
- Gregory P. Asner, Roberta E. Martin, David E. Knapp, Raul Tupayachi, Christopher Anderson, Loreli Carranza, Paola Martinez, Mona Houcheime, Felipe Sinca, and Parker Weiss. Spectroscopy of canopy chemicals in humid tropical forests. *Remote Sensing of Environment*, 115(12):3587–3598, 2011. doi: 10.1016/j.rse.2011.08.020. URL <http://dx.doi.org/10.1016/j.rse.2011.08.020>.

- AxisMaps. Bivariate Choropleth, 2020. URL <https://www.axismaps.com//guide/bivariate-choropleth>.
- Olena Babak and Clayton V. Deutsch. Statistical approach to inverse distance interpolation. *Stochastic Environmental Research and Risk Assessment*, 23(5):543–553, 2009. ISSN 14363240. doi: 10.1007/s00477-008-0226-6.
- C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996. ISSN 0098-3500. doi: 10.1145/235815.235821. URL <https://doi.org/10.1145/235815.235821>.
- Christopher W. Bater. Timelapse.
- Christopher W. Bater, Nicholas C. Coops, Michael A. Wulder, Thomas Hilker, Scott E. Nielsen, Greg Mcdermid, and Gordon B. Stenhouse. Using digital time-lapse cameras to monitor species-specific understorey and overstorey phenology in support of wildlife habitat assessment. *Environmental Monitoring and Assessment*, 180(1-4):1–13, 2010. ISSN 0167-6369. doi: <http://dx.doi.org/10.1007/s10661-010-1768-x>. URL <https://www.proquest.com/docview/880344355/abstract/2BB3B3C0B51E4C20PQ/1>.
- Ant Beck. Specular and Diffuse Reflection, December 2012a. URL [https://commons.wikimedia.org/wiki/File:Specular\\_And\\_Diffuse\\_Reflection.svg](https://commons.wikimedia.org/wiki/File:Specular_And_Diffuse_Reflection.svg).
- Anthony Beck. Airborne Laser Scanning Discrete Echo and Full Waveform signal comparison, 2012b. URL [https://commons.wikimedia.org/wiki/File:Airborne\\_Laser\\_Scanning\\_Discrete\\_Echo\\_and\\_Full\\_Waveform\\_signal\\_comparison.svg](https://commons.wikimedia.org/wiki/File:Airborne_Laser_Scanning_Discrete_Echo_and_Full_Waveform_signal_comparison.svg). Publication Title: Wikimedia Commons.
- Reinette Biggs, Alta de Vos, Rika Preiser, Hayley Clements, Kristine Maciejewski, and Maja Schlüter, editors. *The Routledge Handbook of Research Methods for Social-Ecological Systems*. Routledge, London, July 2021. ISBN 978-1-00-302133-9. doi: 10.4324/9781003021339. URL <https://doi.org/10.4324/9781003021339>.
- Michelle A. Bouchard and US Geological Survey. Example of the Landsat 8 Collection 2 products | U.S. Geological Survey, 2013. URL <https://www.usgs.gov/media/images/example-landsat-8-collection-2-products>.
- Peter Burrough and Rachael McDonnell. Principles of Geographical Information Systems by Peter A. *Oxford University Press*, 54(2):56–57, 1998. ISSN 0028-8144. doi: 10.1111/j.1745-7939.1998.tb02089.x.
- Jürgen Böhner and Th Selige. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. 2006.
- Canadian Space Agency and NASA. Newfoundland – Earth as seen by David Saint-Jacques, 2019. URL <https://www.asc-csa.gc.ca/eng/multimedia/search/Image/Watch/13330>. Last Modified: 2020-06-12.

- James R. Carr and Nai hsien Mao. A general form of probability kriging for estimation of the indicator and uniform transforms. *Mathematical Geology*, 25(4):425–438, 1993. ISSN 08828121. doi: 10.1007/BF00894777.
- Cartographer3d. Concept of LiDAR, 2021. URL [https://commons.wikimedia.org/wiki/File:Concept\\_of\\_LiDAR.svg](https://commons.wikimedia.org/wiki/File:Concept_of_LiDAR.svg). Publication Title: Wikimedia Commons.
- City of Vancouver. LiDAR 2018, a. URL <https://opendata.vancouver.ca/expl ore/dataset/lidar-2018/map/?location=13,49.23783,-123.16721>.
- City of Vancouver. Open Data Portal, b. URL <https://opendata.vancouver.ca/>.
- City of Vancouver, 2009. URL <https://opendata.vancouver.ca/pages/licence/>.
- Vancouver Board of Parks and Recreation City of Vancouver. Street trees, 2012. URL <https://opendata.vancouver.ca/explore/dataset/street-trees/>.
- Emily Clark. *Historical trends of ecosystem services in Canada, 1911-2011*. PhD thesis, McGill University, 2016. URL <https://escholarship.mcgill.ca/concern /theses/pc289m86m?locale=en>.
- I Clark and W V Harper. *Practical Geostatistics 2000*. EcoSSe North America Llc, Ohio, 2007. ISBN 0-9703317-0-3.
- J.K Cliff, A.D. and Ord. Spatial autocorrelation. *Progress in Human Geography*, 19:245–249, 1973.
- Nicholas C. Coops. AMSPEC radiometer.
- Nicholas C. Coops, Thomas Hilker, Michael A. Wulder, Benoît St-Onge, Glenn Newnham, Anders Siggins, and J. A. Trofymow. Estimating canopy structure of Douglas-fir forest stands from discrete-return LiDAR. *Trees - Structure and Function*, 21(3):295–310, 2007. doi: 10.1007/s00468-006-0119-6.
- J W Coulston and G A Reams. The effect of blurred plot coordinates on interpolating forest biomass : a case study 3041 Cornwallis Road 3041 Cornwallis Road. m.
- N. A. C. Cressie. Statistics for Spatial Data, Revised Edition., 1994.
- Eric P. Crist and Richard C. Cicone. A Physically-Based Transformation of Thematic Mapper Data—The TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote Sensing*, GE-22(3):256–263, 1984. ISSN 1558-0644. doi: 10.1109/TGRS.1984.350619.
- Paul J. Curran. Remote sensing of foliar chemistry. *Remote Sensing of Environment*, 30(3):271–278, 1989. doi: 10.1016/0034-4257(89)90069-2.
- Cyp and Kjell André. Polyhedra, 2005. URL <https://commons.wikimedia.org/wiki/Polyhedra>.
- Michele Dalponte and David A. Coomes. Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data. *Methods in*

- Ecology and Evolution*, 7(10):1236–1245, 2016. ISSN 2041210X. doi: 10.1111/2041-210X.12575.
- Bethan Davies. Accuracy and precision in chronostratigraphy, 2020a. URL <https://www.antarcticglaciers.org/glacial-geology/dating-glacial-sediments-2/precision-and-accuracy-glacial-geology/>.
- Bethan Davies. Precision versus accuracy, 2020b. URL <https://www.antarcticglaciers.org/glacial-geology/dating-glacial-sediments-2/precision-and-accuracy-glacial-geology/>.
- Ali Akbar Daya and Hadi Bejari. A comparative study between simple kriging and ordinary kriging for estimating and modeling the Cu concentration in Chehlkureh deposit, SE Iran. *Arabian Journal of Geosciences*, 8(8):6003–6020, 2015. ISSN 18667538. doi: 10.1007/s12517-014-1618-1. ISBN: 1251701416181.
- Lindsay N. Deel, Brenden E. McNeil, Philip G. Curtis, Shawn P. Serbin, Aditya Singh, Keith N. Eshleman, and Philip A. Townsend. Relationship of a Landsat cumulative disturbance index to canopy nitrogen and forest structure. *Remote Sensing of Environment*, 118:40–49, 2012. doi: 10.1016/j.rse.2011.10.026. URL <http://dx.doi.org/10.1016/j.rse.2011.10.026>.
- B. N. Delaunay. Sur la Sphere Vide. In *Izvestia Akademii Nauk SSSR, VII Seria, Otdelenie Matematicheskii i Estestvennyka Nauk*, number 7, pages 793–800. 1934.
- Eric Deluca and Dudley Bonsal. Design and Symbolization. In *Mapping, Society, and Technology*. University of Minnesota Libraries Publishing, 2017. URL <http://open.lib.umn.edu/mapping/chapter/4-design-and-symbolization/>. Book Title: Mapping, Society, and Technology Publisher: University of Minnesota Libraries Publishing.
- Ben den Engelsen. Photo by Ben den Engelsen on Unsplash, 2020. URL <https://unsplash.com/photos/UFwW97AP0LI>.
- C Deutsch and A Journel. GSLIB: Geostatistical Software Library and User’s Guide. 1993.
- David DiBiase. Census Data and Thematic Maps. In *Nature of Geographic Information*. 2014. URL <https://opentextbc.ca/natureofgeographicinformation/chapter/1-overview-2/>.
- DrBob. A negative lens, 2006a. URL <https://commons.wikimedia.org/wiki/File:Lens1b.svg>.
- DrBob. A positive lens, 2006b. URL <https://commons.wikimedia.org/wiki/File:Lens1.svg>.
- Earth Resources Observation and Science Center. USGS EROS Archive - Digital Elevation - Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global,

2018. URL <https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1>.
- David A. Eash, Kimberlee K. Barnes, Padraig S. O'Shea, and Brian K. Gelder. Stream-channel and watershed delineations and basin-characteristic measurements using lidar elevation data for small drainage basins within the Des Moines Lobe landform region in Iowa. *U.S. Geological Survey Scientific Investigations Report 2017-5108*, 2018. ISSN 2328-0328. doi: <https://doi.org/10.3133/sir20175108>. URL <https://pubs.er.usgs.gov/publication/sir20175108>.
- Martin E. Elmer. *Symbol Considerations for Bivariate Thematic Mapping*. Thesis, University of Wisconsin, 2012. URL <https://minds.wisconsin.edu/handle/1793/67887>. Accepted: 2014-01-17T22:34:42Z.
- Graham D. Finlayson and Steven D. Hordley. Color constancy at a pixel. *JOSA A*, 18(2):253–264, February 2001. doi: 10.1364/JOSAA.18.000253. URL <https://www.osapublishing.org/josaa/abstract.cfm?uri=josaa-18-2-253>.
- T Fisher and C MacDonald. An Overview of the Canada Geographic Information System (CGIS). In *Proceedings of the International Symposium on Cartography and Computing: Applications in Health and Environment*, volume 1, pages 610–615, Reston, Virginia, 1979. URL <https://cartogis.org/docs/proceedings/archive/auto-carto-4-vol-1/index3.html>.
- Ministry of Forests and Range Forest Practices Branch. Silviculture Information Submission Guide, 2005. URL [https://www.for.gov.bc.ca/hfp/publications/00026/fs708-14-appendix\\_d.htm#ad\\_02](https://www.for.gov.bc.ca/hfp/publications/00026/fs708-14-appendix_d.htm#ad_02).
- Eric Gaba. Cartographic style guide, 2014. URL [https://en.wikipedia.org/w/index.php?title=Typography\\_\(cartography\)&oldid=1063714862](https://en.wikipedia.org/w/index.php?title=Typography_(cartography)&oldid=1063714862). Page Version ID: 1063714862.
- John C Gallant and Trevor I Dowling. A multiresolution index of valley bottom flatness for mapping depositional areas. *Water resources research*, 39(12), 2003. Publisher: Wiley Online Library.
- R . C . Geary. The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, 5(3):115–127, 1954. URL <https://www.jstor.org/stable/2986645>.
- Arthur Getis and Jared Aldstadt. Constructing the spatial weights matrix using a local statistic. *Advances in Spatial Science*, 61(2):147–163, 2010. ISSN 21979375. doi: 10.1007/978-3-642-01976-0\_11.
- Arthur Getis, Jesus Mur, Bernard Fingleton, and Maria Plotnikova. Spatial Econometrics and Spatial Statistics Spatial Econometrics and Spatial Statistics Arthur Getis , Jesús Mur Lacambra and Henry G . Zoller. (May 2014), 2004.
- Tristan R H Goodbody, Nicholas C Coops, Joan E Luther, Piotr Tompalski, Christopher Mulverhill, Catherine Frizzle, Richard Fournier, Shane Furze, and Sam Herniman. Airborne laser scanning for quantifying criteria and

- indicators of sustainable forest management in Canada. 985(July):972–985, 2021.
- Pierre Goovaerts. Kriging and Semivariogram Deconvolution in the Presence of Irregular Geographical Units. *Mathematical Geology*, 40(1):101–128, 2008. ISSN 15378276. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf>.
- Statistics Canada Government of Canada. Focus on Geography Series, 2016 Census, February 2017. URL <https://www12.statcan.gc.ca/census-recensement/2016/as-sa/fogs-spg/Facts-cma-eng.cfm?GC=933&GK=CMA&LANG=Eng&TOPIC=1>. Last Modified: 2017-02-08.
- R. Graham. An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set. *Information Processing Letters*, 1(4):132–133, 1972. doi: 10.1016/0020-0190(72)90045-2.
- graph-tool. Centrality measures. URL <https://graph-tool.skewed.de/static/doc/centrality.html>.
- Walter Gray. Resources for Tomorrow. *The Globe and Mail*, page 7, January 1962.
- David D. Greenlee. Raster and Vector Processing for Scanned Linework. *Photogrammetric Engineering and Remote Sensing*, 53(10):1383–1387, 1987.
- M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160):850–853, 2013. doi: 10.1126/science.1244693. URL <https://www.science.org/doi/10.1126/science.1244693>.
- K. J. Harker, L. Arnold, Ira J. Sutherland, and Sarah E. Gergel. Perspectives from landscape ecology can improve environmental impact assessment. *FACETS*, 6(1):358–378, 2021.
- Txomin Hermosilla, Michael A. Wulder, Joanne C. White, Nicholas C. Coops, and Geordie W. Hobart. Disturbance-Informed Annual Land Cover Classification Maps of Canada’s Forested Ecosystems for a 29-Year Landsat Time Series. *Canadian Journal of Remote Sensing*, 44(1):67–87, 2018. doi: 10.1080/07038992.2018.1437719. URL <https://doi.org/10.1080/07038992.2018.1437719>.
- M. H. K. Hesselbarth, M. Scaini, K. A. With, K. Wiegand, and J. Nowosad. landscapemetrics: an open-source R tool to calculate landscape metrics, 2019. URL <https://r-spatialecology.github.io/landscapemetrics/>.
- Bill Hillier, Alan Penn, Julienne Hanson, T Grajewski, and J Xu. Natural Movement: Or, Configuration and Attraction in Urban Pedestrian Movement. *Environment and Planning B: Planning and Design*, 20(1):29–66, 1993. ISSN 0265-8135, 1472-3417. doi: 10.1068/b200029.

- Carrie R. Howell, Wei Su, Ariann F. Nassel, April A. Agne, and Andrea L. Cherrington. Area based stratified random sampling using geospatial technology in a community-based survey. *BMC Public Health*, 20(1):1–9, 2020. ISSN 14712458. doi: 10.1186/s12889-020-09793-0. Publisher: BMC Public Health.
- Chang Huang, Ba Duy Nguyen, Shiqiang Zhang, Senmao Cao, and Wolfgang Wagner. A comparison of terrain indices toward their ability in assisting surface water mapping from Sentinel-1 data. *ISPRS International Journal of Geo-Information*, 6(5):140, 2017. Publisher: Multidisciplinary Digital Publishing Institute.
- Juha Hyppä and Mikko Inkinen. Detecting and estimating attributes for single trees using laser scanner. *The Photogrammetric Journal of Finland*, 16(2):27–42, 1999.
- Glenn Iceton. "Many Families of Unseen Indians": Trapline Registration and Understandings of Aboriginal Title in the BC-Yukon Borderlands. *BC Studies*, (201):67–91,175, 2019. ISSN 00052949. URL <https://www.proquest.com/docview/2265666649/abstract/665EA1CCEFE34A59PQ/1>.
- ICSM. Types of Maps, 2021. URL <https://www.icsm.gov.au/education/fundamentals-mapping/types-maps>.
- Deep Inamdar, Margaret Kalacska, George Leblanc, and J Pablo Arroyo Mora. Characterizing and Mitigating Sensor Generated Spatial Correlations in Airborne Hyperspectral Imaging Data. 2020. doi: 10.3390/rs12040641.
- InductiveLoad and NASA. EM Spectrum Properties, October 2007. URL [https://commons.wikimedia.org/wiki/File:EM\\_Spectrum\\_Properties\\_edit.svg](https://commons.wikimedia.org/wiki/File:EM_Spectrum_Properties_edit.svg).
- Ulrike Ingram. Map Elements and Design Principles. In *Introduction to Cartography*. University System of Georgia, 2021. URL <https://alg.manifoldap.org/read/introduction-to-cartography/section/b9662d0a-4926-4947-922d-a67a2cae0eec#2>.
- Edward Isaaks and R.Mohan Srivastava. *An Introduction to Applied Geostatistics.*, volume 17. Oxford University Press, New York, New York, 1st edition, 1989. doi: [https://doi.org/10.1016/0098-3004\(91\)90055-I](https://doi.org/10.1016/0098-3004(91)90055-I). URL <https://www.sciencedirect.com/science/article/pii/009830049190055I>.
- Michel Jaboyedoff, Thierry Oppikofer, Antonio Abellán, Marc Henri Derron, Alex Loyer, Richard Metzger, and Andrea Pedrazzini. Use of LIDAR in landslide investigations: A review. *Natural Hazards*, 61(1):5–28, 2012. ISSN 0921030X. doi: 10.1007/s11069-010-9634-2. ISBN: 1106901096.
- Michelle Jackson, Sarah Gergel, and Kathy Martin. Citizen science and field survey observations provide comparable results for mapping Vancouver Island White-tailed Ptarmigan (*Lagopus leucura saxatilis*) distributions. *Biological Conservation*, 181(6), January 2015. doi: 10.1016/j.biocon.2014.11.010. URL <http://dx.doi.org/10.1016/j.biocon.2014.11.010>.

- Geoffrey M. Jacquez. Spatial statistics when locations are uncertain. *Geographic Information Sciences*, 5(2):77–87, 1999. ISSN 10824006. doi: 10.1080/10824006.2009909480517.
- Marek K. Jakubowski, Wenkai Li, Qinghua Guo, and Maggi Kelly. Delineating individual trees from lidar data: A comparison of vector- and raster-based segmentation approaches. *Remote Sensing*, 5(9):4163–4186, 2013. ISSN 20724292. doi: 10.3390/rs5094163. ISBN: 10.3390/rs5094163.
- R. A. Jarvis. On the identification of the convex hull of a finite set of points in the plane. *Information Processing Letters*, 2(1):18–21, 1973. ISSN 0020-0190. doi: 10.1016/0020-0190(73)90020-3. URL <https://www.sciencedirect.com/science/article/pii/0020019073900203>.
- A. G. Journel. Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology*, 15(3):445–468, 1983. ISSN 00205958. doi: 10.1007/BF01031292.
- Jrvz. An image showing all three axes, 2010. URL [https://commons.wikimedia.org/wiki/File:Yaw\\_Axis\\_Corrected.svg](https://commons.wikimedia.org/wiki/File:Yaw_Axis_Corrected.svg).
- Maxat Kassen. A promising phenomenon of open data: A case study of the Chicago open data project. *Government Information Quarterly*, 30(4):508–513, October 2013. doi: 10.1016/j.giq.2013.05.012. URL <http://dx.doi.org/10.1016/j.giq.2013.05.012>.
- C. H. Key and N. C. Benson. Landscape Assessment: Ground measure of severity, the Composite Burn Index; and Remote sensing of severity, the Normalized Burn Ratio. Technical Report RMRS-GTR-164-CD: LA 1-51, USDA Forest Service, Rocky Mountain Research Station, Ogden, UT, 2006. URL <https://pubs.er.usgs.gov/publication/2002085>.
- Alfred Korzybski. *Science and sanity: an introduction to non-Aristotelian systems and general semantics*. International Non-Aristotelian Library Pub. Co., 1st ed. edition, 1933. Open Library ID: OL24876034M.
- D.G. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.
- Laura La Bella. *Not enough to drink: pollution, drought, and tainted water supplies*. The Rosen Publishing Group, New York, 2009.
- Kathryn A. Lee, Jonathan R. Lee, and Patrick Bell. A review of Citizen Science within the Earth Sciences: potential benefits and obstacles. *Proceedings of the Geologists' Association*, 131(6):605–617, December 2020. ISSN 0016-7878. doi: 10.1016/j.pgeola.2020.07.010. URL <https://www.sciencedirect.com/science/article/pii/S0016787820300730>.
- Gordon Leggett. Google Maps car and camera used for collecting Street View

- data in Steveston, BC Canada, April 2014. URL [https://commons.wikimedia.org/wiki/File:2014-04-29\\_Google\\_Maps\\_Streetview\\_car.jpg](https://commons.wikimedia.org/wiki/File:2014-04-29_Google_Maps_Streetview_car.jpg).
- Library and Archives Canada. Athabasca Glacier from below Wilcox peak, 1917. URL <https://central.bac-lac.gc.ca/.redirect?app=fonandcol&id=4939507&lang=eng>. Last Modified: Date modified (YYYY-MM-DD) / Date de modification (AAAA-MM-JJ).
- Kevin Lim, Paul Treitz, Michael Wulder, Benoît St-Ongé, and Martin Flood. LiDAR remote sensing of forest structure. *Progress in Physical Geography*, 27 (1):88–106, 2003. ISSN 03091333. doi: 10.1191/0309133303pp360ra.
- Patrick J. Little, John S. Richardson, and Younes Alila. Channel and landscape dynamics in the alluvial forest mosaic of the Carmanah River valley, British Columbia, Canada. *Geomorphology*, 202:86–100, November 2013. ISSN 0169-555X. doi: 10.1016/j.geomorph.2013.04.006. URL <https://ui.adsabs.harvard.edu/abs/2013Geomo.202...86L>.
- M. J. López García and V. Caselles. Mapping burns and natural reforestation using thematic Mapper data. *Geocarto International*, 6(1):31–37, 1991. ISSN 1010-6049. doi: 10.1080/10106049109354290. URL <https://doi.org/10.1080/10106049109354290>.
- Kyle Maguire. Mt. Assiniboine, day, a.
- Kyle Maguire. Mt. Assiniboine, sunset, b.
- Aliaksei Makarau, Rudolf Richter, Rupert Müller, and Peter Reinartz. Haze Detection and Removal in Remotely Sensed Multispectral Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(9):5895–5905, September 2014. doi: 10.1109/TGRS.2013.2293662.
- E.W. Manning and J.D. McCuaig. An Overview of the Significance of Urban Centres to Canada's Quality Agricultural Land. Technical Report Report Number 11, Fisheries and Environment Canada Lands Directorate, Ottawa, 1977.
- Steven Manson. Hue, 2013a. URL <https://open.lib.umn.edu/mapping/chapter/4-design-and-symbolization/#footnote-399-12>.
- Steven Manson. Symbolization, 2013b. URL <https://open.lib.umn.edu/mapping/chapter/4-design-and-symbolization/#footnote-399-6>.
- Steven Manson. Value, 2013c. URL <https://open.lib.umn.edu/mapping/chapter/4-design-and-symbolization/#footnote-399-13>.
- Steven Manson. Map composition. In *Mapping, Society, and Technology*. University of Minnesota Libraries Publishing, 2017. URL <https://open.lib.umn.edu/mapping/chapter/4-design-and-symbolization/>. Book Title: Mapping, Society, and Technology Publisher: University of Minnesota Libraries Publishing.

- Steven Manson and Laura Matson. Maps, Society, and Technology. In *Mapping, Society, and Technology*. University of Minnesota Libraries Publishing, 2017. URL <https://open.lib.umn.edu/mapping/chapter/1-maps-society-and-technology/>.
- Marek9134. LiDAR-i lend, 2012. URL [https://commons.wikimedia.org/wiki/File:LiDAR-i\\_lend.gif](https://commons.wikimedia.org/wiki/File:LiDAR-i_lend.gif). Publication Title: Wikimedia Commons.
- Danny Marks, Jeff Dozier, and James Frew. Automated basin delineation from digital elevation data. *Geo-processing*, 2:299–311, 1984.
- Nicholas Martino. Spatial Network Analysis, 2020. URL <https://ubc-library-rc.github.io/qgis-walkability/>. Publication Title: Spatial Network Analysis.
- G Matheron. *Traité de géostatistique appliquée*. Number v. 1 in Mémoires. Éditions Technip, 1962. URL <https://books.google.ca/books?id=88YKAQAAQAJ>.
- G. Matheron. Intrinsic Random Functions and Their Applications. *Advances in Applied Probability*, 5(3):439–468, 1973. ISSN 00018678. doi: 10.1017/S0001867800039379.
- Pietro Mattivi, Francesca Franci, Alessandro Lambertini, and Gabriele Bitelli. TWI computation: a comparison of different open source GISs. *Open Geospatial Data, Software and Standards*, 4(1):1–12, 2019. Publisher: SpringerOpen.
- MAXAR. Precision3D Data Suite. URL <https://www.maxar.com/products/precision3d-data-suite>.
- Mc Clapurhands. LiDARUSA Snoopy 120 LiDAR, 2019. URL [https://commons.wikimedia.org/wiki/File:Yellowscan\\_LIDAR\\_on\\_OnyxStar\\_FOX-C8\\_HD.jpg](https://commons.wikimedia.org/wiki/File:Yellowscan_LIDAR_on_OnyxStar_FOX-C8_HD.jpg). Publication Title: Wikimedia Commons.
- Loren McClenahan, Andrew B. Cooper, Matthew G. McKenzie, and Joshua A. Drew. The Importance of Surprising Results and Best Practices in Historical Ecology. *BioScience*, 65(9):932–939, September 2015. ISSN 0006-3568. doi: 10.1093/biosci/biv100. URL <https://doi.org/10.1093/biosci/biv100>.
- James McCrorie. ARDA: An Experiment in Development Planning. Technical Report Special Study 2, Canadian Council on Rural Development, Ottawa, 1969.
- Bruce McCune and Dylan Keon. Equations for potential annual direct incident radiation and heat load. *Journal of Vegetation Science*, 13:603–606, August 2002. doi: 10.1111/j.1654-1103.2002.tb02087.x. URL [https://www.researchgate.net/publication/280685772\\_Equations\\_for\\_potential\\_annual\\_direct\\_incident\\_radiation\\_and\\_heat\\_load](https://www.researchgate.net/publication/280685772_Equations_for_potential_annual_direct_incident_radiation_and_heat_load).
- R. E. McRoberts, E. O. Tomppo, and R. L. Czaplewski. Sampling designs for national forest assessments: Knowledge reference for national forest assessments. *Food and Agricultural Organisation of the UN (FAO)*, pages 23–40,

2014. URL <http://www.fao.org/%0Aforestry/44859-02cf95ef26dfdc86c6be2720f8b938a8.pdf>. Rome.FAO.
- Qingmin Meng, Chris J. Ciesewski, Mike R. Strub, and Bruce E. Borders. Spatial regression modeling of tree height-diameter relationships. *Canadian Journal of Forest Research*, 39(12):2283–2293, 2009. ISSN 00455067. doi: 10.1139/X09-136.
- David Monniaux. Lidar P1270901.jpg, 2007. URL [https://commons.wikimedia.org/wiki/File:Lidar\\_P1270901.jpg](https://commons.wikimedia.org/wiki/File:Lidar_P1270901.jpg). Publication Title: Wikimedia Commons.
- David Monniaux, cmglee, and jimhtatshawdotca. Illustration of the method Eratosthenes used to calculate the circumference of the Earth, 2005. URL [https://commons.wikimedia.org/wiki/File:Eratosthenes\\_measure\\_of\\_Earth\\_circumference.svg](https://commons.wikimedia.org/wiki/File:Eratosthenes_measure_of_Earth_circumference.svg).
- P. Moran. Notes on Continuous Stochastic Phenomenon. *Biometrika*, 37(1): 17–23, 1950.
- J. L. Morgan, S. Gergel, Collin Ankerson, S. Tomscha, and Ira J. Sutherland. Historical Aerial Photography for Landscape Analysis. In *Learning Landscape Ecology*, pages 21–40. Springer, New York, NY, 2017. ISBN 978-1-4939-6374-4. URL [https://doi.org/10.1007/978-1-4939-6374-4\\_2](https://doi.org/10.1007/978-1-4939-6374-4_2).
- Salem Morsy, Ahmed Shaker, and Ahmed El-Rabbany. Multispectral lidar data for land cover classification of urban areas. *Sensors (Switzerland)*, 17(5), 2017. ISSN 14248220. doi: 10.3390/s17050958.
- Mountain Legacy Project. Modern Athabasca Glacier, 2011. URL <http://mountainlegacy.ca/>.
- Nadar. The Arc de Triomphe and the Grand Boulevards, Paris, from a Balloon, 1868. URL [https://commons.wikimedia.org/wiki/File:Nadar\\_triumph\\_1868.jpg](https://commons.wikimedia.org/wiki/File:Nadar_triumph_1868.jpg).
- NASA. Atmospheric Transmission, a. URL <https://landsat.gsfc.nasa.gov/satellites/landsat-8/>.
- NASA. GOES-1 Satellite, b. URL [keeptrack.space](http://keeptrack.space).
- NASA and Tom Neumann. IceSat-2; Ice, Cloud, AND LAND ELEVATION SA LITE-2, 2021. URL <https://icesat-2.gsfc.nasa.gov/>.
- NASA and University of Maryland. GEDI Ecosystem LiDAR, 2021. URL <https://gedi.umd.edu/>.
- Native Land. Native Land Digital. URL <https://native-land.ca/>.
- Natural Resources Canada. Earth Observation Data Management System, a. URL <https://www.eodms-sgdot.nrcan-rncan.gc.ca/index-en.html>.
- Natural Resources Canada. Natural Resources Canada, b. URL <https://www.nrcan.gc.ca/home>.

- Natural Resources Canada. Canadian Digital Elevation Model, 1945-2011, 2015. URL <https://open.canada.ca/data/en/dataset/7f245e4d-76c2-4caa-951a-45d1d2051333>.
- Canadian Forest Service Natural Resources Canada. Douglas-Fir and Western Hemlock, December 2013. URL <https://tidef.nrcan.gc.ca/en/trees/factsheet/119>. Last Modified: 2015-07-24.
- Ross Nelson. How did we get here? An early history of forestry lidar. *Canadian Journal of Remote Sensing*, 39(SUPPL.1), 2013. ISSN 07038992. doi: 10.5589/m13-011.
- NEON Education. A key component of a lidar system is a GPS, 2014. URL <https://www.flickr.com/photos/128087132@N06/15408803390>. Publication Title: Flickr.
- Norman Leon Nicholson and Paul Comtois. Growing seasons, 1950. URL <https://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~323192~90092294:-24-Growing-seasons->.
- Norman Leon Nicholson and Paul Comtois. Bedrock geology, 1956. URL <https://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~323176~90092279:-16-Bedrock-geology->.
- Norman Leon Nicholson and Paul Comtois. Comparison of scales., 1958. URL <https://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~323156~90092341:-6--Comparison-of-scales->.
- NRCAN. Canadian Digital Elevation Model | Earth Engine Data Catalog, 2021. URL [https://developers.google.com/earth-engine/datasets/catalog/NRCan\\_CDEM](https://developers.google.com/earth-engine/datasets/catalog/NRCan_CDEM).
- Erik Næsset. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment*, 80(1):88–99, 2002. ISSN 00344257. doi: 10.1016/S0034-4257(01)00290-5. ISBN: 0034-4257.
- Ministry of Agriculture and Food. Food Land Guidelines: A policy statement of the government of Ontario on planning for agriculture. Technical report, Government of Ontario, Toronto, 1978.
- Department of Regional Economic Expansion. The Canada Land Inventory: Objectives, Scope, and Organization. Technical Report Report 1, Queen's Printer for Canada, Ottawa, 1965.
- Dominion Bureau of Statistics. Number and area of occupied farms. Technical report, Ottawa, 1952. URL <https://archive.org/details/1951981951m31952engfra/mode/2up>.
- Dominion Bureau of Statistics. Volume I: Population General Characteristics. Technical report, Ottawa, 1953. URL <https://archive.org/details/1951981951FV11953engfra/page/n17/mode/2up>.

- Atsuyuki Okabe, Barry Boots, and Kokichi Sugihara. Nearest neighbourhood operations with generalized Voronoi diagrams: a review. *International Journal of Geographical Information Systems*, 8(1):43–71, 1994. ISSN 0269-3798. doi: 10.1080/02693799408901986. URL <https://doi.org/10.1080/02693799408901986>.
- OpenStreetMap. URL <https://www.openstreetmap.org/copyright>.
- Oregon Department of Transportation. 3D mobile mapping unit, 2016. URL <https://www.flickr.com/photos/oregondot/24784201084/in/photostream/>. Publication Title: Flickr.
- Oxford Languages. Phenomena. URL <https://languages.oup.com/google-dictionary-en/>.
- J. Anthony Parker, Robert V. Kenyon, and Donald E. Troxel. Comparison of Interpolating Methods for Image Resampling. *IEEE Transactions on Medical Imaging*, 2(1):31–39, March 1983. ISSN 1558-254X. doi: 10.1109/TMI.1983.4307610. URL <https://ieeexplore.ieee.org/document/4307610>. Conference Name: IEEE Transactions on Medical Imaging.
- Edzer J. Pebesma, Richard N.M. Duin, and Peter A. Burrough. Mapping sea bird densities over the North Sea: Spatially aggregated estimates and temporal changes. *Environmetrics*, 16(6):573–587, 2005. ISSN 11804009. doi: 10.1002/env.723.
- Paul Pickell, Nicholas Coops, Colin Ferster, Christopher Bater, Karen Blouin, Mike Flannigan, and Jinkai Zhang. An early warning system to forecast the close of the spring burning window from satellite-observed greenness. *Scientific Reports*, 7(1), October 2017. doi: 10.1038/s41598-017-14730-0.
- Peter Potapov, Xinyuan Li, Andres Hernandez-Serna, Alexandra Tyukavina, Matthew C. Hansen, Anil Kommareddy, Amy Pickens, Svetlana Turubanova, Hao Tang, Carlos Edibaldo Silva, John Armston, Ralph Dubayah, J. Bryan Blair, and Michelle Hofton. Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sensing of Environment*, 253, February 2021. ISSN 0034-4257. doi: 10.1016/j.rse.2020.112165. URL <https://www.sciencedirect.com/science/article/pii/S0034425720305381>.
- Province of BC. Provincial Change Monitoring Inventory (CMI) and Young Stand Monitoring (YSM) Sampling Framework. (May), 2018. URL [https://www2.gov.bc.ca/assets/gov/farming-natural-resources-and-industry/forestry/stewardship/forest-analysis-inventory/ground-sample-inventories/provincial-monitoring/provincial\\_cmi\\_and\\_ysm\\_sampling\\_framework\\_20180616.pdf](https://www2.gov.bc.ca/assets/gov/farming-natural-resources-and-industry/forestry/stewardship/forest-analysis-inventory/ground-sample-inventories/provincial-monitoring/provincial_cmi_and_ysm_sampling_framework_20180616.pdf).
- C. Qin, A.-X. Zhu, T. Pei, B. Li, C. Zhou, and L. Yang. An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. *International Journal of Geographical Information Science*, 21(4):

- 443–458, April 2007. ISSN 1365-8816. doi: 10.1080/13658810601073240. URL <https://doi.org/10.1080/13658810601073240>.
- Quartl. English: Illustration of the cartesian product A x B of two sets A={x,y,z} and B={1,2,3}., October 2012. URL [https://commons.wikimedia.org/wiki/File:Cartesian\\_Product\\_qtl1.svg](https://commons.wikimedia.org/wiki/File:Cartesian_Product_qtl1.svg).
- Romain Quey. Neper. URL <https://github.com/rquey/neper>.
- Paul Ramsey. 23. Linear Referencing — Introduction to PostGIS, 2012. URL [http://postgis.net/workshops/postgis-intro/linear\\_referencing.html](http://postgis.net/workshops/postgis-intro/linear_referencing.html).
- R.K. Raney, A.P. Luscombe, E.J. Langham, and S. Ahmed. RADARSAT (SAR imaging). *Proceedings of the IEEE*, 79(6):839–849, June 1991. doi: 10.1109/5.90162. URL <http://dx.doi.org/10.1109/5.90162>.
- Anthony C. Robinson. Layout and Symbolization. In *Maps and the Geospatial Revolution*. The Pennsylvania State University, 2020. URL [https://www.e-education.psu.edu/maps/l5\\_p3.html](https://www.e-education.psu.edu/maps/l5_p3.html).
- Robert A. Rohde. Solar Radiation Spectrum, October 2008. URL [https://commons.wikimedia.org/wiki/File:Solar\\_Spectrum.png](https://commons.wikimedia.org/wiki/File:Solar_Spectrum.png).
- Richard E. Rossi, David J. Mulla, André G. Journel, and Eldon H. Franz. Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs*, 62(2):277–314, 1992. ISSN 00129615. doi: 10.2307/2937096.
- J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering. Monitoring vegetation systems in the Great Plains with ERTS. In *NASA. Goddard Space Flight Center 3d ERTS-1 Symp.*, volume 1, 1974. URL <https://ntrs.nasa.gov/citations/19740022614>.
- Jean Romain Roussel, David Auty, Nicholas C. Coops, Piotr Tompalski, Tristan R.H. Goodbody, Andrew Sánchez Meador, Jean François Bourdon, Florian de Boissieu, and Alexis Achim. lidR: An R package for analysis of Airborne Laser Scanning (ALS) data. *Remote Sensing of Environment*, 251 (August):112061, 2020. ISSN 00344257. doi: 10.1016/j.rse.2020.112061. URL <https://doi.org/10.1016/j.rse.2020.112061>. Publisher: Elsevier.
- Jean-Romain Roussel, Tristan R.H. Goodbody, and Piotr Tompalski. lidR Book, 2021. URL <https://r-lidar.github.io/lidRbook/>.
- Shannon1. Longest Rivers of Canada, 2017. URL [https://commons.wikimedia.org/wiki/File:Longest\\_Rivers\\_of\\_Canada.png](https://commons.wikimedia.org/wiki/File:Longest_Rivers_of_Canada.png).
- D Shepard. Two- dimensional interpolation function for irregularly- spaced data. *Proc 23rd Nat Conf*, pages 517–524, 1968.
- Ronald L Shreve. Statistical law of stream numbers. *The Journal of Geology*, 74(1):17–37, 1966. Publisher: University of Chicago Press.

- June Skeeter, Andreas Christen, and Greg H.R. Henry. Controls on Carbon Dioxide and Methane fluxes from a Low-Center Polygonal Peatland in the Mackenzie River Delta. *Arctic Science*, February 2022. doi: 10.1139/AS-2021-0034. URL <https://cdnsciencepub.com/doi/abs/10.1139/AS-2021-0034>. Publisher: NRC Research Press.
- John Parr Snyder. *Map projections—A working manual*, volume 1395. US Government Printing Office, 1987.
- Special Committee on Land Use in Canada. Consolidation of the proceedings and considerations of the committee from its inception on January 30, 1957 to the end of the first Session 25th Parliament, February 6th, 1963 on Land Use in Canada, December 1963. (Chair: Arthur M. Pearson).
- Statistics Canada. Statistics Canada. URL <https://www.statcan.gc.ca/en/start>.
- Census of Population Statistics Canada. Thematic maps – Age and sex – Percentage of the population aged 65 and over in 2016, by census tract (CT), 2016. URL <https://www12.statcan.gc.ca/census-recensement/2016/geo/map-carte/ref/thematic-thematiques/as/map-eng.cfm?type=5&UID=305>. Last Modified: 2017-05-03.
- Census of Population Statistics Canada. Illustrated Glossary | Thematic map, November 2017. URL <https://www150.statcan.gc.ca/n1/pub/92-195-x/2016/001/other-autre/theme/theme-eng.htm>. Last Modified: 2017-11-15.
- Government of Canada Statistics Canada. Canadian Index of Multiple Deprivation: Dataset, July 2019. URL <https://www150.statcan.gc.ca/n1/pub/45-20-0001/452000012019001-eng.htm>. Last Modified: 2019-07-12.
- R G Steel and J Torrie. Principles and procedures of statistics: a biometrical approach (2nd ed). 1980.
- Ayal Stein, Eran Geva, and Jihad El-Sana. CudaHull: Fast parallel 3D convex hull on the GPU. *Computers & Graphics*, 36(4):265–271, 2012. ISSN 0097-8493. doi: 10.1016/j.cag.2012.02.012. URL <https://www.sciencedirect.com/science/article/pii/S0097849312000350>.
- Arthur N Strahler. Quantitative analysis of watershed geomorphology. *Eos, Transactions American Geophysical Union*, 38(6):913–920, 1957. Publisher: Wiley Online Library.
- K. Suryowati, R. D. Bekti, and A. Faradila. A Comparison of Weights Matrices on Computation of Dengue Spatial Autocorrelation. *IOP Conference Series: Materials Science and Engineering*, 335(1), 2018. ISSN 1757899X. doi: 10.1088/1757-899X/335/1/012052.
- Anu Swatantran, Hao Tang, Terence Barrett, Phil Decola, and Ralph Dubayah. Rapid, high-resolution forest structure and terrain mapping over large ar-

- eas using single photon lidar. *Scientific Reports*, 6(June):1–12, 2016. ISSN 20452322. doi: 10.1038/srep28277. Publisher: Nature Publishing Group.
- Systems Innovation. Graph Theory Overview, April 2015a. URL <https://youtu.be/82zlRaRUsaY>.
- Systems Innovation. Network Diffusion & Contagion, April 2015b. URL <https://youtu.be/bTXUJQhEqL0>.
- David G. Tarboton. A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water Resources Research*, 33(2):309–319, 1997. URL <https://doi.org/10.1029/96WR03137>.
- Abbas Tashakkori and Charles Teddlie. SAGE Handbook of Mixed Methods in Social & Behavioral Research, 2010. URL <https://methods.sagepub.com/book/sage-handbook-of-mixed-methods-social-behavioral-research-2e>.
- Charles Teddlie and Fen Yu. Mixed Methods Sampling: A Typology With Examples. *Journal of Mixed Methods Research*, 1(1):77–100, 2007. ISSN 15586901. doi: 10.1177/2345678906292430.
- The Center for International Forestry Research (CIFOR). LiDAR machine, 2014. URL <https://www.flickr.com/photos/cifor/36019786455/>. Publication Title: Flickr.
- The Good Doctor Fry. Diagram of scattering of sunlight to make a yellow/red appearing sun and blue sky, January 2011. URL [https://commons.wikimedia.org/wiki/File:Rayleigh\\_mie\\_fry3a.jpg](https://commons.wikimedia.org/wiki/File:Rayleigh_mie_fry3a.jpg).
- The Senate of Canada. Minutes of the Proceedings of the Senate of Canada, January 1957.
- S K Thompson. *Sampling*. CourseSmart. Wiley, 2012. ISBN 978-1-118-16294-1. URL <https://books.google.ca/books?id=-sFtXLIdDiIC>.
- Shanley D. Thompson, Trisalyn A. Nelson, Joanne C. White, and Michael A. Wulder. Mapping Dominant Tree Species over Large Forested Areas Using Landsat Best-Available-Pixel Image Composites. *Canadian Journal of Remote Sensing*, 41(3):203–218, May 2015. doi: 10.1080/07038992.2015.1065708. URL <https://doi.org/10.1080/07038992.2015.1065708>.
- Jency Titus and Sebastian Geroge. A Comparison Study On Different Interpolation Methods Based On Satellite Images. *International Journal of Engineering Research & Technology*, 2(6):82–85, 2013. ISSN 2278-0181. URL [www.ijert.org](http://www.ijert.org).
- Author W R Tobler. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46:234–240, 1970. ISSN 0036-8075. doi: 10.1126/science.ns-13.332.462.
- Matthew J. Tomlinson, Sarah E. Gergel, Timothy J. Beechie, and Michelle M. McClure. Long-term changes in river—floodplain dynamics: implications for

- salmonid habitat in the Interior Columbia Basin, USA. *Ecological Applications*, 21(5):1643–1658, 2011. ISSN 1051-0761. URL <https://www.jstor.org/stable/23023107>.
- Roger Tomlinson. *An introduction to the Geo-information system of the Canada Land Inventory*. Department of Forestry and Rural Development, Ottawa, 1967.
- Roger Tomlinson. *The application of electronic computing methods and techniques to the storage, compilation, and assessment of mapped data*. Dissertation, University of London, London, 1974.
- Piotr Tompalski, Joanne C. White, Nicholas C. Coops, and Michael A. Wulder. Demonstrating the transferability of forest inventory attribute models derived using airborne laser scanning data. *Remote Sensing of Environment*, 227 (September 2018):110–124, 2019. ISSN 00344257. doi: 10.1016/j.rse.2019.04.006. URL <https://doi.org/10.1016/j.rse.2019.04.006>. Publisher: Elsevier.
- Stephanie A. Tomscha, Ira J. Sutherland, Delphine Renard, Sarah E. Gergel, JEANINE M. Rhemtulla, Elena M. Bennett, Lori D. Daniels, Ian M. S. Eddy, and Emily E. Clark. A Guide to Historical Data Sets for Reconstructing Ecosystem Service Change over Time. *BioScience*, 66(9):747–762, 2016. ISSN 0006-3568. URL <https://www.jstor.org/stable/90007657>.
- TransLink. 2020 RapidBus & B-Line Network, 2020. URL [https://www.translink.ca/resources/translink/plansandprojects/busprojects/rapidbus/rapidbus\\_network\\_map](https://www.translink.ca/resources/translink/plansandprojects/busprojects/rapidbus/rapidbus_network_map).
- Panagiotis Tziachris, Eirini Metaxa, Frantzin Papadopoulos, and Maria Papadopoulou. Spatial modelling and prediction assessment of soil iron using Kriging interpolation with pH as auxiliary information. *ISPRS International Journal of Geo-Information*, 6(9), 2017. ISSN 22209964. doi: 10.3390/ijgi6090283.
- UF Geomatics - Fort Lauderdale. LiDAR Remote Sensing Part 3: Data Analysis, 2016a. URL <https://youtu.be/wpW1cx7WMXQ>.
- UF Geomatics - Fort Lauderdale. LiDAR Remote Sensing Part 2: Systems & Data Collection, 2016b. URL <https://youtu.be/jS0II-ZMSSo>.
- UF Geomatics - Fort Lauderdale. LiDAR Remote Sensing Part 1: Principles, 2016c. URL <https://youtu.be/T5hFrS57LGo>.
- Kent State University. The Cuyahoga River Watershed. In *Proceedings of a symposium commemorating the dedication of Cunningham Hall*, Cleveland, November 1968. Kent State University.
- University of Texas Libraries. Intro to Georeferencing, 2021. URL <https://guides.lib.utexas.edu/georeference-raster-data/intro-to-georeferencing>.
- USGS. Copyrights and Credits | U.S. Geological Survey. URL <https://www.usgs.gov/information-policies-and-instructions/copyrights-and-credits>.

- Susan L. Ustin and Michael E. Schaepman. Imaging spectroscopy special issue. *Remote Sensing of Environment*, 113(SUPPL. 1):S1, 2009. doi: 10.1016/j.rse.2008.09.018. URL <http://dx.doi.org/10.1016/j.rse.2008.09.018>.
- Todd Vorenkamp. How Focus Works, 2015. URL <https://www.bhphotovideo.com/explora/photography/tips-and-solutions/how-focus-works>.
- Hans Wackernagel. *Multivariate Geostatistics: An Introduction with Applications*. Springer, 3rd edition, 2002. ISBN 978-3-662-05294-5. doi: 10.1007/978-3-662-05294-5.
- Ran Wang, John A. Gamon, Jeannine Cavender-Bares, Philip A. Townsend, and Arthur I. Zygierbaum. The spatial sensitivity of the spectral diversity-biodiversity relationship: An experimental test in a prairie grassland. *Ecological Applications*, 28(2):541–556, 2018. doi: 10.1002/eap.1669.
- Martin Wegmann. Hemispherical photo in the Bavarian Forest, October 2011. URL [https://commons.wikimedia.org/wiki/File:Hemispherical\\_photo1.jpg](https://commons.wikimedia.org/wiki/File:Hemispherical_photo1.jpg).
- Carolyn Weiss, Patricia Cillis, and Neil Rothwell. The Population Ecumene of Canada: Exploring the Past and Present. *Geography Working Paper Series*, 2008. URL <https://www150.statcan.gc.ca/n1/pub/92f0138m/92f0138m2008003-eng.htm>.
- Joanne C. White, Michael A. Wulder, Andrés Varhola, Mikko Vastaranta, Nicholas C. Coops, Bruce D Cook, Doug Pitt, and Murray Woods. A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. Information Report FI-X-010. Technical report, Canadian Forest Service, Canadian Wood Fibre Centre, Pacific Forestry Centre, Victoria, BC, Canada, 2013. ISBN: 9781100223858 ISSN: 00157546.
- Wikibooks. Communication Networks/Network Topologies, 2018. URL [https://en.wikibooks.org/wiki/Communication\\_Networks/Network\\_Topologies](https://en.wikibooks.org/wiki/Communication_Networks/Network_Topologies).
- Wikimedia. Delaunay Triangulation. *Wikipedia*, July 2021a.
- Wikimedia. PageRank. *Wikipedia*, July 2021b.
- Phil Wilkes, Simon D. Jones, Lola Suarez, Andrew Haywood, William Woodgate, Mariela Soto-Berelov, Andrew Mellor, and Andrew K. Skidmore. Understanding the Effects of ALS Pulse Density for Metric Retrieval across Diverse Forest Types. *Photogrammetric Engineering & Remote Sensing*, 81(8):625–635, 2015. ISSN 00991112. doi: 10.14358/PERS.81.8.625. URL <http://openurl.ingenta.com/content/xref?genre=article&issn=0099-1112&volume=81&issue=8&spage=625>. ISBN: 1872-8332 (Electronic)\n0169-5002 (Linking).
- Natalie Wolchover. How Far Can the Human Eye See?, May 2012. URL <https://www.livescience.com/33895-human-eye.html>.

- Worldmapper. No Water Access per capita, 2015. URL <https://worldmapper.org/maps/no-water-access-per-capita/>.
- Ikuho Yamada. Thiessen Polygons. *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, pages 1–6, 2016. doi: 10.1002/9781118786352.wbieg0157. ISBN: 9781118786352.
- S. R. Yates, A. W. Warrick, and D. E. Myers. Disjunctive Kriging: 2. Examples. *Water Resources Research*, 22(5):623–630, 1986. ISSN 19447973. doi: 10.1029/WR022i005p00623.
- Hamdi A. Zurqani, Christopher J. Post, Elena A. Mikhailova, Michael P. Cope, Jeffery S. Allen, and Blake A. Lytle. Evaluating the integrity of forested riparian buffers over a large area using LiDAR data and Google Earth Engine. *Scientific Reports*, 10(1):1–16, 2020. ISSN 20452322. doi: 10.1038/s41598-020-69743-z. URL <https://doi.org/10.1038/s41598-020-69743-z>. ISBN: 0123456789 Publisher: Nature Publishing Group UK.
- Joanne Zwinkels. Encyclopedia of Color Science and Technology. *Encyclopedia of Color Science and Technology*, pages 1–8, 2020. doi: 10.1007/978-3-642-27851-8.