Characterization of Volcanic Terrains Using Lidar Reflectivity: A Statistical Approach

Michael Barber

Follow this and additional works at: https://knowledge.library.iup.edu/etd

Part of the Applied Mathematics Commons

Recommended Citation
https://knowledge.library.iup.edu/etd/1593

This Thesis is brought to you for free and open access by Knowledge Repository @ IUP. It has been accepted for inclusion in Theses and Dissertations (All) by an authorized administrator of Knowledge Repository @ IUP. For more information, please contact cclouser@iup.edu, sara.parne@iup.edu.
CHARACTERIZATION OF VOLCANIC TERRAINS USING LIDAR REFLECTIVITY:
A STATISTICAL APPROACH

A Thesis
Submitted to the School of Graduate Studies and Research
in Partial Fulfillment of the
Requirements for the Degree
Master of Science

Michael Barber
Indiana University of Pennsylvania
May 2018
We hereby approve the thesis of

Michael Barber

Candidate for the degree of Master of Science

Frederick A. Adkins, Ph.D.
Professor of Mathematics, Advisor

Nicholas D. Deardorff, Ph.D.
Assistant Professor of Geoscience

Yongtao Cao, Ph.D.
Assistant Professor of Mathematics

ACCEPTED

Randy L. Martin, Ph.D.
Dean
School of Graduate Studies and Research
Title: Characterization of Volcanic Terrains using Lidar Reflectivity: A Statistical Approach

Author: Michael Barber

Thesis Chair: Dr. Frederick A. Adkins

Thesis Committee Members:  
Dr. Nicholas D. Deardorff  
Dr. Yongtao Cao

In recent decades, lidar has revolutionized topographic mapping of the Earth and planets through the use of digital elevation models (DEMs). However, the return amplitudes of the reflected laser pulses, typically collected as part of a lidar dataset, have seldom been used as a means of identifying and characterizing volcanic surface features such as lava flows, rafted tephra and agglutinate, and pyroclastic deposits consisting of tephra and ashfall. Here, we find an effective process for remotely characterizing volcanic terrains using a simple but rigorous cluster analysis of lidar return amplitudes and DEM data to define the parameters for a self-organizing mapping routine. The data used for this study, collected from the Northwest Rift Zone on Newberry Volcano in central Oregon, has been accurately georeferenced, providing 3 foot horizontal and 4.5 centimeter vertical resolution. In addition, the return amplitude values were recorded with a horizontal resolution of 1.5 feet. An appropriate number of terrain categories is found by applying an incremental within-cluster sum of squares algorithm to generate clusters from random subsets of lidar DEM and reflectivity data chosen from the study area. From these results, a silhouette analysis determines the optimum number of clusters to be used as a necessary input parameter for the categorization of each data cell by means of a self-organizing mapping function. These results, confirmed by comparison with field work conducted at Mokst Butte and its associated lava flows, is then applied to other, less accessible volcanic terrains. The resulting false-color imagery allows precise identification of volcanic morphologies that are otherwise unrecognizable in remote sensing data such as lidar, InSAR, and orthorectified color photography, and in regions where traditional field work is difficult or unfeasible.
Dedicated to my mother.

This research was funded in part by the National Science Foundation Scholarships in Science, Technology, Engineering, and Mathematics (S-STEM) program under Award No. DUE 1259860, and by the Geological Society of America under the Stephen Pollock Undergraduate Research Program.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Content</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.1 Background</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.2 Previous Research</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.3 Present Work</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>STUDY AREA</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>2.1 Newberry Volcano and The Northwest Rift Zone</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>2.2 Mokst Butte</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2.3 Field Observations and Sample Analysis</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>DATA</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>3.1 Lidar</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>3.2 Variables</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>COMPUTATION AND ANALYSIS</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>4.1 Objective</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>4.2 Clustering</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>4.3 Silhouette Analysis</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>4.4 Interpretation of Silhouette Analysis</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>RESULTS</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>5.1 Self-Organizing Map</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>5.2 Interpretation of Results</td>
<td>23</td>
</tr>
<tr>
<td>6</td>
<td>DISCUSSION</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>6.1 Application</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>6.2 Conclusions</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>REFERENCES</td>
<td>34</td>
</tr>
<tr>
<td>8</td>
<td>APPENDIX</td>
<td>37</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Laboratory Reflectances and Corresponding Expected Lidar Reflectivities for</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Materials Found in the Mokst Butte Study Area.</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Summary Statistics of Lidar Reflectivity by Class.</td>
<td>25</td>
</tr>
<tr>
<td>5.3</td>
<td>Mean Reflectivity, Inferred Terrain Type, and Color Assignment by Class.</td>
<td>26</td>
</tr>
<tr>
<td>6.1</td>
<td>Summary Statistics of Lidar Reflectivity by Class.</td>
<td>31</td>
</tr>
<tr>
<td>6.2</td>
<td>Mean Reflectivity, Inferred Terrain Type, and Color Assignment by Class.</td>
<td>32</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Location of Newberry Volcano in Central Oregon.</td>
<td>4</td>
</tr>
<tr>
<td>2.2</td>
<td>Mokst Butte Study Area, Shown in Lidar DEM Derived Hillshade Image.</td>
<td>6</td>
</tr>
<tr>
<td>2.3</td>
<td>Mokst Butte Study Area, Shown in Color Orthophoto Image.</td>
<td>7</td>
</tr>
<tr>
<td>2.4</td>
<td>Mokst Butte Cone Area.</td>
<td>8</td>
</tr>
<tr>
<td>2.5</td>
<td>Mokst Butte Proximal Lava Flows.</td>
<td>8</td>
</tr>
<tr>
<td>3.1</td>
<td>Moving Box of 9 Cells Used by the Total Curvature Tool.</td>
<td>11</td>
</tr>
<tr>
<td>3.2</td>
<td>Histograms of Curvature, Height, and Reflectivity Data.</td>
<td>14</td>
</tr>
<tr>
<td>3.3</td>
<td>Box-Cox Normality Plots of Curvature and Height Data.</td>
<td>14</td>
</tr>
<tr>
<td>3.4</td>
<td>Histograms of Transformed Curvature, Height, and Reflectivity Data.</td>
<td>15</td>
</tr>
<tr>
<td>4.1</td>
<td>Silhouette Plots Showing Results of Four Experiments.</td>
<td>20</td>
</tr>
<tr>
<td>4.2</td>
<td>Box Placed Over the Region of Interest in the Silhouette Plot for Experiment 1.</td>
<td>20</td>
</tr>
<tr>
<td>4.3</td>
<td>Silhouette Plots Showing Averaged Results of Each Experiment Normalized to Naturally Decreasing Values.</td>
<td>21</td>
</tr>
<tr>
<td>5.1</td>
<td>Raw Output from Iso Cluster Unsupervised Classification in ArcGIS.</td>
<td>23</td>
</tr>
<tr>
<td>5.2</td>
<td>Boxplot of Reflectivity Values by Class for Iso Unsupervised Classification of Mokst Butte Study Area.</td>
<td>25</td>
</tr>
<tr>
<td>5.3</td>
<td>Interpreted False-Color Image of Iso Classification from Mokst Butte Study Area.</td>
<td>27</td>
</tr>
<tr>
<td>5.4</td>
<td>Annotated False-Color Image of Iso Classification from Mokst Butte Study Area.</td>
<td>27</td>
</tr>
<tr>
<td>5.5</td>
<td>Mokst Butte Cone Area False-Color Image.</td>
<td>28</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.6</td>
<td>Mokst Butte Proximal Lava Flows False-Color Image.</td>
<td>28</td>
</tr>
<tr>
<td>6.1</td>
<td>Raw Output from Iso Cluster Unsupervised Classification in ArcGIS for the Forest Road Flow.</td>
<td>30</td>
</tr>
<tr>
<td>6.2</td>
<td>Boxplot of Reflectivity Values by Class for Forest Road Flow.</td>
<td>31</td>
</tr>
<tr>
<td>6.3</td>
<td>Interpreted False-Color Image of Iso Classification of Forest Road Flow.</td>
<td>32</td>
</tr>
</tbody>
</table>
Lidar ("light radar") was developed shortly after the invention of the laser in the early 1960s, and had its first well-known use in accurately measuring the distance from the Earth to the Moon [26]. The principle of lidar is straightforward: a pulse of laser light emitted by the device is reflected from a target object and is detected upon its return. The two-way travel time of this pulse is used to accurately determine the distance from the device to the reflecting object [24]. The development of high pulse rates and improved portability, in combination with recently developed very accurate GPS positioning methods, have allowed lidar to revolutionize the acquisition of high quality topographic data from the Earth, Moon, and other bodies of the Solar System [2, 31, 32, 36]. To obtain topographical data from the surface of the Earth, the laser scanning equipment is often mounted on aircraft whose position is known in relation to portable, ground-based GPS stations. Such airborne lidar provides data with horizontal accuracies on a sub-meter to meter scale (limited by spot-size as a function of pulse travel distance), and vertical resolutions on the order of centimeters [34]. This spatial topographic data is now widely used for mapping, but the laser reflection intensity data, a relative measure of the amount of energy reflected from the target, which is also usually recorded as part of a lidar dataset, has seldom been used to analyze the spectral and textural features of landforms [17, 20].

1.2 Previous Research

Because lidar reflectivity is largely a function of the reflective character of the target at the wavelength of the laser pulse, it has been shown that classification and identification of target materials is possible using lidar reflection intensity data [3, 25]. Lidar reflectivity data has been widely used in the identification and classification of trees, crops and other vegetation [1, 12, 16, 22], and to a lesser degree as a tool to classify and identify sedimentary and volcanic rocks in the terrains surrounding the volcanoes of Stromboli and Mt. Etna, in
Italy [3,10,20]. The purposes of these studies varied widely, and as a result, the methods employed in the collection, processing, and analysis of the data did as well. For example, Fornaciai et al [11] undertook a lidar intensity analysis of Stromboli volcano as an assessment of the potential to use lidar intensity for the characterization of fundamental volcanic products such as pyroclastic deposits, epiclastic sediments, and lava flows, Favalli et al [7] used lidar intensity data as an aid in mapping a single channel-fed lava flow field on Mt. Etna, and Mazzarini et al [20] used lidar intensity as a means of identifying the chronology of multiple lava flows spanning approximately 400 years on Mt. Etna.

1.3 Present Work

All of the volcanic terrain studies noted above relied heavily on field observations for interpretation of the lidar intensity and other data, and the subsequent analysis and discussion of results were treated accordingly. In effect, the lidar intensity data was used primarily as an aid to direct field work or to help confirm interpretations derived from other sources.

A different approach was taken by Kim et al [15], who used a k-medoids clustering method, with lidar intensity and height above ground of lidar returns as input data, to classify tree species without relying on field observations. This, combined with a silhouette analysis, indicated two groups as the optimum clustering level: one cluster contained deciduous trees and the other cluster contained evergreen trees. However, once these two groups were identified, no further subgroups or individual identifications were made directly from the data, and individual tree “species” within each cluster were identified only by field observation (a mix of species and genus names were ultimately given).

In the present work, the author extends the underlying approach of Kim et al to classify and identify objects from lidar DEM and reflectivity data using a clustering algorithm and silhouette analysis, but several important differences should be noted. The study area consists of volcanic terrains (primarily lava flows and pyroclastic deposits) as well as forested
and vegetated areas. Simple random samples are used to reduce the size of the dataset to a manageable level for the required computations. This allows analysis and interpretation of lidar data covering a much larger area (∼10 km$^2$ vs. ∼1 km$^2$). The silhouette analysis is used to determine optimum numbers of clusters to be used as necessary (and difficult to obtain) inputs for self-organizing mapping and classification algorithms. A much greater range and depth of classifications will be shown to be possible and robust. And, finally, the interpreted data will be used to produce useful and easily interpreted map products.
CHAPTER 2

STUDY AREA

2.1 Newberry Volcano and The Northwest Rift Zone

Newberry Volcano, east of the High Cascades in central Oregon, is one of the largest Quaternary volcanoes in the conterminous United States. It is a broad shieldlike landform that covers an area approximately the size of the state of Rhode Island, and its highest point, on the southwest rim of the summit caldera, rises 7,989 feet above sea level. It comprises the products of thousands of eruptions during the past 600,000 years, at least 25 of which occurred during the last 10,000 years (Holocene Epoch), and is thus considered an active volcano, with the potential to threaten the activities and safety of approximately 200,000 people who live and work within 60 miles of it. The most recent eruptive events include the Northwest Rift Zone eruptions ~7,000 years ago, and the Big Obsidian Flow ~1,300 years ago [18].

Figure 2.1: Location of Newberry Volcano in Central Oregon. Inset on left shows Northwest Rift Zone lava flows (black) extending north-northwest from Newberry caldera toward the city of Bend, Oregon.
The Northwest Rift Zone (NWRZ) lies on the northwest flank of Newberry Volcano, extending nearly 20 miles from the Newberry summit caldera toward the city of Bend, Oregon (figure 2.1). The NWRZ includes multiple lava flows, scoria cones, fissure vents, and pyroclastic deposits produced by a series of related eruptions which are constrained in age by radiocarbon dating to 6610 to 7240 years before present (calibrated calendar age) [18]. This is more recent than the Mt. Mazama explosion of \( \sim 7650 \) years ago that covered the entire Newberry region with a heavy ashfall, and as a result, the NWRZ features, overlying the Mazama ash, are very well exposed. This, combined with reasonably good access via a network of U.S. Forest Service roads, make the study of a wide range of volcanic morphologies possible.

### 2.2 Mokst Butte

There are over 400 parasitic scoria cones of varying ages on the flanks of Newberry Volcano, and many were named using words from the Pacific Northwest trade language of Chinook Jargon [14, 18]. In this language, “Mokst” means “two.” The eruptions at Mokst Butte, accounting for nearly half of the total eruptive volume of the NWRZ events, emanated from several vents, producing a scoria cone some 700 feet high and lava flows extending nearly 6 miles to the north-northwest. At least two times in this series of eruptions, increased magma flux caused the southwest portion of the cone to collapse, resulting in its current prominence of 500 feet. Large quantities of cone material, consisting mostly of oxidized scoria and agglutinate, was subsequently rafted away on the flowing lava to the full extent of the flow field. Other volcanic features, such as lava flows not containing rafted cone material, lava which has been moved since primary emplacement (either naturally or by human activities), and airfall tephra were also produced.
2.3 Field Observations and Sample Analysis

The specific area serving as the basis for the work presented here, shown in figure 2.2, covers an area of 10 km$^2$ and comprises the vents and proximal lava flows associated with the Mokst Butte eruptions.

Field work conducted by the author in July 2015 and July 2016, and by prior workers [14,18,21], comprises direct observations, measurement, and sample analysis. This field data provides a general overview and understanding of the Mokst Butte eruptions, and allows positive identification of the following features, located in figures 2.4 and 2.5:

a) Scoria (black in color), in-situ inside lower portion of Mokst Butte cone.

b) Scoria (black in color), on inner slope of Mokst Butte cone, moved from primary emplacement by mass wasting.

c) Oxidized scoria (red in color), in-situ on top rim of Mokst Butte cone.

d) Oxidized scoria (red in color), on upper inner slope of Mokst Butte cone, moved from primary emplacement by mass wasting.
e) Weathered basaltic andesite lava, in-situ on early flow lobe approximately 1 km west of Mokst Butte cone. Sample collected by author confirmed by XRF analysis to contain 57.07 wt% SiO$_2$, 5.41 wt% Na$_2$O/K$_2$O.

f) Basaltic andesite lava, same as above, disturbed by timber harvesting operations.

g) Weathered andesite lava, in-situ on late flow lobe approximately 0.5 km north-northwest of Mokst Butte Cone. Sample collected by author confirmed by XRF analysis to contain 58.58 wt% SiO$_2$, 5.86 wt% Na$_2$O/K$_2$O.

h) Rafted cone material (oxidized scoria, bombs, and agglutinate), on late flow lobe approximately 1 km west of Mokst Butte cone.

i) Lapilli- and ash-sized tephra from the Mt. Mazama and Mokst Butte eruptions, near Mokst Butte lava flow approximately 1 km west of Mokst Butte cone.

j) Conifer forest / individual trees, on upper east side of Mokst Butte cone.

k) Shrubland (low vegetation primarily consisting of sagebrush), near Mokst Butte lava flow approximately 1 km west of Mokst Butte cone.

l) Dead/dried vegetation (pine needles), accumulated by wind along Mokst Butte flow front approximately 1 km west of Mokst Butte cone.

Figure 2.3: Mokst Butte Study Area, Shown in Color Orthophoto Image. Boxes locate closer views labeled in figures 2.4 and 2.5 below.
Figure 2.4: Mokst Butte Cone Area. Features labeled: a) scoria, black in color, in-situ; b) scoria, black in color, moved/tumbled by mass wasting; c) oxidized scoria, red in color, in-situ; d) oxidized scoria, red in color, moved/tumbled by mass wasting; g) weathered andesite, in-situ; j) conifer forest.

Figure 2.5: Mokst Butte Proximal Lava Flows. Features labeled: i) ash- and lapilli-sized tephra; k) shrubland; l) pine needles; h) rafted cone material; e) weathered basaltic andesite lava, in-situ; f) basaltic andesite lava, disturbed by logging operations.
CHAPTER 3

DATA

3.1 Lidar

Lidar data was collected between May 28th and July 27th, 2010 under contract with Watershed Sciences as part of the Newberry CVO Lidar Project Delivery 2 for the Oregon Department of Geology and Mineral Resources. This delivery includes data in the form of grids, trajectory files, intensity images, Lidar ASCII Standard (LAS) point files, ground point density rasters, Real Time Kinematic (RTK) survey data, a shapefile of the delivery area, and the lidar delivery report [19].

The lidar survey utilized Leica ALS50 Phase II and ALS60 sensors mounted in multiple Cessna Caravan 208B aircraft. The Leica systems, operating at a wavelength of 1,064 nm (near infrared), were set to acquire at least 105,000 laser pulses per second (i.e. 105 kHz pulse rate) and flown at 900 and 1300 meters above ground level (AGL), capturing a scan angle of ±14 degrees from nadir. These settings are developed to yield points with an average density of at least 8 points per square meter over terrestrial surfaces. To solve for laser point position, aircraft position was described as x, y and z, and measured twice per second (2 Hz) by an onboard differential GPS unit. Aircraft attitude was measured 200 times per second (200 Hz) as pitch, roll and yaw (heading) from an onboard inertial measurement unit (IMU). The GPS survey utilized a Trimble GPS receiver model R7 with a Zephyr Geodetic antenna and ground plane for static control points. A Trimble GPS R8 unit was primarily used for gathering RTK locations but was also used as a static receiver. For RTK data, the collector began recording after remaining stationary for 5 seconds then calculated the pseudo range position from at least three epochs with the relative error under 1.5cm horizontal and 2cm vertical. All GPS measurements were made with dual frequency L1-L2 receivers with carrier-phase correction [33].
3.2 Variables

The lidar data used for this study are the “bare earth” and “highest hit” digital elevation models (DEMs), each with 3 ft horizontal and 4.5 cm vertical resolutions, and the lidar return intensity raster, with 1.5 ft per pixel resolution, all part of the Newberry CVO Lidar Project described above, and obtained by download from the Oregon Department of Geology and Mineral Resources website.

First of the three measurements used to identify and characterize specific terrain types is lidar reflectivity, or intensity. The lidar return intensity raster consists of one return amplitude value per cell (pixel), collected from the first return or “highest hit” from each laser pulse. These amplitudes are recorded as integers ranging from a low of 0 to a high of 254 (8-bit data) on a linear scale.

The second measurement used is surface curvature, serving as a proxy for surface roughness on a meter-scale. Surface roughness of lava flows, which is a function of lava block size, is a commonly recognized indicator of the general nature of the lava. Typically, higher silica content (e.g., andesite vs. basalt), results in larger block size and greater surface roughness [30]. This is evident throughout the NWRZ, where andesite lava flows such as the distal flow lobe at Mokst Butte are dramatically more rugged than basalt lava flows such as the forest road flow and the Lava Cast Forest flow. In addition to roughness differences caused by different bulk mineral compositions, lava flows almost always show higher surface roughness than the terrain surrounding them in lidar bare earth models. This is clearly visible in figure 2.2, where the lava flows are easily discerned from the smoother surrounding areas.

To quantify surface curvature, the “Total Curvature” option in the DEM Surface Tools for ArcGIS was used to calculate surface curvature from the bare earth DEM [13]. This tool uses a $3 \times 3$ moving box of 9 cells, as shown in figure 3.1, to compute a surface curvature value for the center cell. The box is then shifted one cell, and the process is repeated until
a new raster with the same horizontal resolution as the original DEM is created, consisting of a curvature value for each pixel.

There are two broad approaches commonly used to calculate curvature from raster digital elevation models: the “Evans” approach [6] and the “Zevenbergen & Thorne” approach [35]. ArcGIS Spatial Analyst uses the Zevenbergen & Thorne approach, while DEM Surface Tools uses the Evans approach. The difference between the two is that the Evans approach fits a curve to the 9 elevation points using a 6-parameter polynomial which is only a best fit to the 9 points and does not necessarily go through all of them, while the Zevenbergen & Thorne approach uses a higher-order 9-parameter polynomial that does go through all 9 points. Florinsky [9] argues in favor of the Evans approach because it filters out small random errors in the original DEM data. Florinsky [8] also shows mathematically that the Evans method is more precise than the Zevenbergen & Thorne method. Schmidt et al. [29] argue that quadratic-based algorithms like those used in the Evans approach are more stable than the partial quartic algorithms of Zevenbergen & Thorne.
Total curvature is computed as follows. Let \( x \) and \( y \) represent the geographic coordinates and let \( z \) represent the elevation of each data cell. Now, define \( r, s, \) and \( t \) as follows:

\[
\begin{align*}
    r &= \frac{\partial^2 z}{\partial x^2}, \quad s = \frac{\partial^2 z}{\partial x \partial y}, \quad \text{and} \quad t = \frac{\partial^2 z}{\partial y^2}.
\end{align*}
\]

Then the total curvature value is given by

\[
\text{Total Curvature} = \left( r^2 + 2s^2 + t^2 \right) \times 100.
\]

Raster elevation data projected in latitude/longitude coordinates are not equally spaced on a square grid, but actually take the form of a trapezoid. To account for this, the numerical computations used to obtain the partial derivatives shown above follow Florinsky’s modification of Evans’ original methods [9], resulting in the following formulae (see figure 3.1 for indexing of variables):

\[
\begin{align*}
    r &= \frac{c^2(z_1 + z_3 - 2z_2) + b^2(z_4 + z_6 - 2z_5) + a^2(z_7 + z_9 - 2z_8)}{a^4 + b^4 + c^4} \\
    s &= \frac{c [a^2(d + e) + b^2 e]}{2 [a^2 c^2 (d + e)^2 + b^2 (a^2 d^2 + c^2 e^2)]} (z_3 - z_1) - b (a^2 d - c^2 e)(z_4 - z_6) + a \left[ c^2(d + e) + b^2 d \right] (z_7 - z_9) \\
    t &= \frac{2 \left[ d (a^4 + b^4 + b^2 e^2) - c^2 e (a^2 - b^2) \right] (z_1 + z_3) - d (a^4 + c^4 + b^2 c^2) + e (a^4 + c^4 + a^2 b^2) \right] (z_4 + z_6) + \\
        \left[ e (b^4 + c^4 + a^2 b^2) + a^2 d (b^2 - c^2) \right] (z_7 + z_9) + d \left[ b^4 (z_2 - 3z_5) + c^4 (3z_2 - z_5) + (a^4 - 2b^2 c^2)(z_2 - z_5) \right] + \\
       \left[ 3de(d + e)(a^4 + b^4 + c^4) \right] + e \left[ a^4 (3z_8 - z_5) + b^4 (z_8 - 3z_5) + (c^4 - 2a^2 b^2)(z_8 - z_5) \right] - 2 \left[ a^2 d (b^2 - c^2) z_8 - c^2 e (a^2 - b^2) z_2 \right].
\end{align*}
\]

These computations produce strictly non-negative curvature values of radians per linear unit. Applied to the present data, the resulting values range from a low of 0 to a high of \( \sim 3 \times 10^4 \) radians per foot. Thus, no distinction is made between positive and negative concavity, which is appropriate when the only consideration is absolute surface roughness.

Note, a curvature value of zero denotes that the 9-cell box is a plane, but does not imply that the plane is horizontal. Multiplying the total curvature values by 100 produces values in
the same general range as the ArcGIS Curvature function, and has no effect on the relative difference in curvature from one cell to another.

The third and final measurement is the difference between the highest hits and bare earth DEMs, denoted as “height” because it primarily represents the height of vegetation (shrubs, trees, etc.), although rough, blocky lava flows could also result in a first and last return of different elevations from the same pulse footprint. This simple subtraction of the elevation of the bare earth DEM from the highest hit DEM produces values ranging from 0 to \( \sim 176 \) feet. This is consistent with the expected minimum derived from relatively smooth, bare surfaces to the maximum that would be obtained from the heights of the tallest trees in the area.

These three measurements all have minimums of zero, but the maximum of \( 3.15 \times 10^4 \) for curvature is much greater than the maximums for reflectivity and height. Also, the distribution of values within each of these measurements is quite dissimilar. Figure 3.2 illustrates this, showing the extreme right-skewed nature of the curvature and height data.

To address the extreme skewness of the curvature and height data, Box-Cox normality plots were created for each dataset. Figure 3.3 shows the plots for curvature (left) and height (right), with values of \( \lambda = 0 \) and \( \lambda = 0.01 \) respectively.

The standard log transformation, where the natural logarithm of all data values is taken, is simply a special case of the Box-Cox transformation with \( \lambda = 0 \). And while it is not necessary that the data used in clustering computations be normally distributed, it is desirable that the values in each dataset be spread over comparable ranges, i.e., within an order of magnitude or so. Thus, the datasets for curvature and height were transformed identically as follows. Let \( y \) represent the original data in a set of \( i \) values to be transformed, and let \( t \) represent the transformed data. Then the transformation \( y_i \rightarrow t_i \) is given by

\[
t_i = \log (y_i + \min\{y > 0\}) - \min \{\log (y_i + \min\{y > 0\})\}.
\]
Figure 3.2: Histograms of Curvature, Height, and Reflectivity Data. Curvature (a), height (b), and reflectivity (c).

Figure 3.3: Box-Cox Normality Plots of Curvature and Height Data. Box-Cox normality plots of curvature (a) and height (b) data, showing $\lambda$-values of 0 and 0.01 respectively.
Histograms of the final datasets used for the subsequent analysis in this study are shown in figure 3.4. The limits of the $x$-axes in these plots are set to the minimum and maximum of each dataset. The lidar reflectivity data had one null value which was removed, as were the corresponding cells in each of the other datasets. All other datapoints are represented in these plots.

![Histograms of Transformed Curvature, Height, and Reflectivity Data.](image)

Figure 3.4: Histograms of Transformed Curvature, Height, and Reflectivity Data. Final curvature (a), height (b), and reflectivity(c) data, as used in subsequent analysis.
CHAPTER 4
COMPUTATION AND ANALYSIS

4.1 Objective

With the ultimate objective of determining an optimal number of classes to be set as a parameter for a self-organizing map algorithm, a combination of partitioning (non-heirarchical clustering) and silhouette analysis is employed. The Mokst Butte study area dataset covers a grid of 5468 rows by 8749 columns, or $\sim 47.8$ million cells, each with three measurements. To facilitate implementation of the computationally expensive methods being used, simple random samples chosen from the dataset reduce the amount of data being evaluated to a feasible yet robust level of 5,000 cells. Each experimental run consists of five trials, each with 5,000 randomly selected, non-repeating cells. Clustering and silhouette analysis is performed for each trial, then the silhouette analyses are averaged to produce the final result of the experimental run. All computations for clustering and silhouette analysis are performed in Matlab using a combination of Mathworks source functions and code written by the author. These codes are shown in Appendix A.

4.2 Clustering

To create clusters of data, Ward’s minimum variance inner squared distance method is employed. Like other clustering algorithms such as UPGMA (unweighted pair-group method using arithmetic averages), SLINK (single linkage clustering), and CLINK (complete linkage clustering), Ward’s method begins with $n$ clusters, each containing one object, and ends with one cluster containing all $n$ objects. At each step, the merger of two clusters that is made is the one which will result in the smallest sum-of-squares variance, or distance metric. Thus, for each clustering step the variances of all possible mergers of two clusters are computed, and the smallest determines the merger chosen. The process is then repeated until all $n$ objects are contained in one cluster [27].
Ward’s algorithm as executed by Matlab finds the distances between all objects in a cluster and the centroid of that cluster, then computes the sum of squares variance as the following distance metric \( d(r, s) \), given by
\[
d(r, s) = \sqrt{\frac{2n_r n_s}{(n_r + n_s)} \| \bar{x}_r - \bar{x}_s \|_2},
\]
where

\( \| \cdot \|_2 \) is the standard 2-norm (Euclidean distance).
\( \bar{x}_r \) and \( \bar{x}_s \) are the centroids of clusters \( r \) and \( s \).
\( n_r \) and \( n_s \) are the number of elements in clusters \( r \) and \( s \).

Note that some formulations of Ward’s method do not multiply \( n_r n_s \) by 2. The Matlab implementation uses this factor to give the distance between two singleton clusters as the Euclidean distance.

### 4.3 Silhouette Analysis

In textbook descriptions of clustering, a group of objects is dendritically divided and sub-divided into small, distinct groups based on a set of attributes that conveniently allow such divisions. However, in actual research applications, computational clustering algorithms are seldom used to group objects that are so easily classified, or for which their interrelationships are even inferred. As a result, the “quality” of the clusters obtained must be evaluated with care.

As Rousseeuw aptly points out, “...clustering methods always come up with \( k \) groups, whatever the data are like.” He goes on to note several factors which must be considered when assessing the quality of a clustering solution, including consideration of the compactness versus separation of the clusters, the relationship of the clusters to any actual structure in the data, and whether or not objects are well- or miss-qualified [28].

17
To address these issues, Rousseeuw introduced what he called “silhouettes” as a graphical aid in determining an appropriate number of clusters [28]. The process of constructing silhouettes is fairly straightforward. Starting with the partition obtained from clustering and the distances between all objects, a value $s(i)$ is computed as follows. For each object $i$ in cluster $A$ containing $n$ objects, the average dissimilarity of $i$ to all other objects in $A$ is given by

$$a(i) = \text{average Euclidean distance between } i \text{ and all other objects in } A.$$ 

Next, some other cluster $C$ not containing $i$ is considered, and the dissimilarity of $i$ to all objects in $C$, given by

$$d(i, C) = \text{average Euclidean distance between } i \text{ and all objects in } C$$

is computed for all clusters $C \neq A$.

Now, the smallest of the values $d(i, C)$ is denoted as

$$b(i) = \min_{C \neq A} \{d(i, C)\}.$$ 

Finally, the silhouette value $s(i)$ is obtained from $a(i)$ and $b(i)$ as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} ; \quad i = 1, \ldots, n.$$ 

From this, it can be seen that $-1 \leq s(i) \leq 1$. Further, it can be seen that as $s(i)$ approaches a value of 1, object $i$ is more appropriately assigned to its cluster, and as $s(i)$ approaches a value of $-1$, object $i$ is more likely assigned to the wrong cluster. For $s(i)$ values close to zero, object $i$ can be considered to be “between” clusters, with no good assignment made.

The benefit of computing values $s(i)$ for each object $i$ is that the average of all $s(i)$ values can be used to consider the relative strength of one clustering solution when compared with another one. The Matlab function `evalclusters` performs the process described above.
over a user-specified range of cluster numbers, and provides a graphical output of silhouette values for each clustering solution in that range.

To cluster and analyze the data for this study, *evalclusters* was used with Ward’s clustering method specified, over a range of 2-20 clusters. A “weighting” parameter was set to treat all clusters identically, regardless of the number of objects in each, and the distance measure used was set to use Euclidean distances for the silhouette analysis, all as detailed above.

### 4.4 Interpretation of Silhouette Analysis

Four experimental runs are presented, each with five trials using 5,000 randomly chosen, non-repeating objects having the three attributes described in Chapter 3. To allow for repeatability, each experiment carries a run identification number representing the random number generator seed that was used for that experiment. Plots of the silhouette values obtained are shown in figure 4.1. For each experiment, the dashed lines represent the silhouette values for each of the five trials, and the solid black line represents the average of the five trials. Though solutions of a very few number of clusters (i.e., 2-4 clusters) are of little interest in the context of this investigation, the full range of solutions tested, from 2 to 20 clusters, is shown to provide an overview of the results.

Figure 4.2 shows a box placed over the region of interest in the silhouette plot for experiment 1. This box, which follows the natural decreasing slope of the silhouette values is expanded in figure 4.3, which shows the averaged data for each experiment normalized to the natural negative slope of decreasing silhouette values. Small and large numbers of clusters have been eliminated, reducing the range of solutions to 6 to 15 clusters. This shows the relative strength of the clustering solutions, and in all four plots the solution of 13 clusters is strong relative to neighboring values. Therefore, 13 clusters is chosen as the “number of classes” parameter in the self-organizing map algorithm.
Figure 4.1: *Silhouette Plots Showing Results of Four Experiments*. Number of clusters shown on x-axis. Silhouette value shown on y-axis. Dashed lines show individual trials, solid lines show average of five trials for each experiment.

Figure 4.2: *Box Placed Over the Region of Interest in the Silhouette Plot for Experiment 1*. This box, which follows the natural decreasing slope of the silhouette values is expanded in figure 4.3.
Figure 4.3: Silhouette Plots Showing Averaged Results of Each Experiment Normalized to Naturally Decreasing Values. Number of clusters shown on $x$-axis. Relative silhouette strength shown on $y$-axis. Dashed line represents natural negative slope of values.
CHAPTER 5

RESULTS

5.1 Self-Organizing Map

The Iso ("Iterative self-organizing") Cluster Unsupervised Classification tool in the ArcGIS Spatial Analyst Toolbox is a powerful function for classifying data. The algorithm employs a combination of iterative clustering and maximum likelihood classification to quickly produce a single-layer raster of classified values from multiple raster inputs. However, this tool like other similar self-organized mapping tools, requires the user to enter the desired number of classes to create from the input data. This non-trivial requirement is generally not well explained in user documentation, and the typical instruction is to choose some number of classes greater than what is likely to be represented in the data. A tutorial on the use of the Iso Unsupervised Classification tool states “How many spectral classes we should set will be dependent on the number of informational classes, the complexity of your area of interest, and your final project. You may need to run the Iso Cluster Unsupervised Classification multiple times using different numbers of spectral classes” [23]. The Esri ArcGIS Desktop “How Iso Cluster Works” web page states “The optimal number of classes to specify is usually unknown. Therefore, it is advised to enter a conservatively high number, analyze the resulting clusters, and rerun the function with a reduced number of classes.” Clearly, specifying an appropriate number of classes based on the inherent structure of the data avoids guesswork, and allows the maximum amount of information to be revealed.

Having determined that 13 classes is an appropriate number for the three datasets described in chapter 3, Iso Cluster Unsupervised Classification is executed with the three dataset rasters selected as inputs and 13 classes specified for the output raster. The raw output is shown in figure 5.1. Each of the 13 classes is present, but randomly colored, and it is difficult to determine what is represented in each class.
5.2 Interpretation of Results

To apply meaningful colors to the classes created, consideration is given to what may be most meaningful in the data. For volcanic terrains, and as discussed in chapter 1, lidar reflectivity can be a key indicator of the nature of volcanic morphologies. To further quantify the relationship between lidar reflectivity and target materials within the study area, several studies which discuss lidar reflectance as related to the mineralogical composition of rocks were consulted [3–5, 10, 20]. Specific reflectance values for the materials relevant to this study were gathered from the USGS Spectral Library. This data repository makes available the reflectance properties of hundreds of materials at several specific wavelengths, including 1.064 µm, which matches the lidar data used in this study. From these sources, a table of material reflectances and corresponding lidar reflectivities can be constructed. The reflectances listed
in the USGS Spectral Library are from laboratory samples measured under controlled conditions, and direct comparison of those figures with actual lidar data collected from aircraft is difficult. However, the reflectance of materials at specific wavelengths is quite consistent, and certain relationships of materials do allow meaningful predictions.

Generally, vegetation, and especially dead vegetation such as the pine needles and deadfall logs which are abundant in the region of this study, are highly reflective in the near-infrared. Somewhat less reflective, and also relevant to this study are iron oxide minerals including the iron oxide hematite, which is the primary coating on red, oxidized scoria, and the iron hydroxide goethite, which is a common alteration mineral associated with iron-rich rocks. Weathered lavas are less reflective at this wavelength than iron oxide minerals, but are still more reflective than fresh, unweathered lava, which has quite low reflectance. An important factor in the reflectivity of lavas is the relative silica content, which creates a continuum of increasing reflectance with increasing silica content. For this study, it is important to note that andesite has a higher silica content and is more reflective than basaltic andesite, which is more reflective than basalt. All of the materials mentioned are fully dispersive (lambertian) reflectors. This is important in eliminating differences in return amplitudes caused by varying angles of incidence. However, specular surfaces (glossy surfaces) also exist in the study area, and unless a specular surface is perfectly normal to the path of the incident laser pulse, virtually none of the pulse will be returned to the receiver. Thus, specular surfaces such as water, glossy leaves, and pine needles produce the lowest reflectivity values.

Table 5.1 shows a list of materials which can be used to characterize the volcanic morphologies and surrounding areas of the Mokst Butte study area. The Mean Reflectance values are from the USGS Spectral Library, and represent the average values of materials falling within the categories listed in the Material column. The Mean Lidar Reflectivity values are derived using dead vegetation as a benchmark, and applying the same factor of 300 to the Mean Reflectances of the other materials. These are the lidar reflectivity values expected from the materials listed.
Table 5.1: Laboratory Reflectances and Corresponding Expected Lidar Reflectivities for Materials Found in the Mokst Butte Study Area.

<table>
<thead>
<tr>
<th>Material</th>
<th>n</th>
<th>Mean Reflectance</th>
<th>Mean Lidar Reflectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>specular surfaces</td>
<td></td>
<td></td>
<td>very low or very high</td>
</tr>
<tr>
<td>fresh basaltic andesite</td>
<td>5</td>
<td>.167</td>
<td>50.1</td>
</tr>
<tr>
<td>weathered basaltic andesite</td>
<td>3</td>
<td>.224</td>
<td>67.2</td>
</tr>
<tr>
<td>weathered andesite</td>
<td>2</td>
<td>.255</td>
<td>76.5</td>
</tr>
<tr>
<td>oxidized scoria</td>
<td>5</td>
<td>.299</td>
<td>89.7</td>
</tr>
<tr>
<td>tephra</td>
<td>3</td>
<td>.409</td>
<td>122.7</td>
</tr>
<tr>
<td>living vegetation</td>
<td>4</td>
<td>.526</td>
<td>157.8</td>
</tr>
<tr>
<td>dead vegetation</td>
<td>7</td>
<td>.656</td>
<td>196.8</td>
</tr>
</tbody>
</table>

Table 5.2: Summary Statistics of Lidar Reflectivity by Class.

<table>
<thead>
<tr>
<th>class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>min</td>
<td>0</td>
<td>1</td>
<td>50</td>
<td>65</td>
<td>75</td>
<td>84</td>
<td>94</td>
<td>103</td>
<td>113</td>
<td>123</td>
<td>139</td>
<td>154</td>
<td>147</td>
</tr>
<tr>
<td>Q1</td>
<td>7</td>
<td>38</td>
<td>56</td>
<td>67</td>
<td>78</td>
<td>88</td>
<td>97</td>
<td>106</td>
<td>116</td>
<td>128</td>
<td>142</td>
<td>158</td>
<td>182</td>
</tr>
<tr>
<td>median</td>
<td>15</td>
<td>44</td>
<td>59</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>99</td>
<td>108</td>
<td>119</td>
<td>131</td>
<td>145</td>
<td>162</td>
<td>188</td>
</tr>
<tr>
<td>Q3</td>
<td>23</td>
<td>48</td>
<td>62</td>
<td>73</td>
<td>83</td>
<td>92</td>
<td>101</td>
<td>111</td>
<td>122</td>
<td>135</td>
<td>149</td>
<td>168</td>
<td>200</td>
</tr>
<tr>
<td>max</td>
<td>33</td>
<td>61</td>
<td>66</td>
<td>75</td>
<td>85</td>
<td>95</td>
<td>105</td>
<td>116</td>
<td>124</td>
<td>139</td>
<td>155</td>
<td>179</td>
<td>254</td>
</tr>
<tr>
<td>mean</td>
<td>14.9</td>
<td>42.6</td>
<td>58.7</td>
<td>70.1</td>
<td>80.2</td>
<td>89.8</td>
<td>98.9</td>
<td>108.5</td>
<td>119.0</td>
<td>131.4</td>
<td>145.6</td>
<td>163.0</td>
<td>192.7</td>
</tr>
<tr>
<td>st dev</td>
<td>8.88</td>
<td>6.23</td>
<td>3.76</td>
<td>2.85</td>
<td>2.59</td>
<td>2.59</td>
<td>2.87</td>
<td>3.16</td>
<td>4.02</td>
<td>4.35</td>
<td>6.33</td>
<td>14.49</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.2: Boxplot of Reflectivity Values by Class for Iso Unsupervised Classification of Mokst Butte Study Area. Note: statistics used for this plot and which are shown in table 5.2 above are reflectivity values only, while classification is based on curvature and height data as well as reflectivity.
The lidar reflectivity values in each of the 13 classes produced by the Iso Unsupervised Classification can now be used to identify the primary materials that make up that class. Table 5.2 and figure 5.2 show reflectivity value summary statistics and side-by-side box plots for each class. It is important to note that the statistics shown in table 5.2 and figure 5.2, and which will be used for the subsequent interpretations are for reflectivity values only, while the class divisions are based on curvature and height data as well as reflectivity. Applying the criteria for lidar reflectivity by material from table 5.1 to each class, false colors can be applied to create a map of terrain types. Table 5.3 shows the classes, mean reflectivity values, inferred terrain type, and false color to be applied.

Table 5.3: Mean Reflectivity, Inferred Terrain Type, and Color Assignment by Class.

<table>
<thead>
<tr>
<th>Class</th>
<th>Mean Reflectivity</th>
<th>Inferred Terrain Type</th>
<th>ArcGIS Color</th>
<th>Color Swatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.9</td>
<td>specular surfaces</td>
<td>yogo blue</td>
<td>🔄</td>
</tr>
<tr>
<td>2</td>
<td>42.6</td>
<td>unweathered lava</td>
<td>ultra blue</td>
<td>🔄</td>
</tr>
<tr>
<td>3</td>
<td>58.7</td>
<td>weathered basaltic andesite</td>
<td>black</td>
<td>💥</td>
</tr>
<tr>
<td>4</td>
<td>70.1</td>
<td>weathered basaltic andesite</td>
<td>black</td>
<td>💥</td>
</tr>
<tr>
<td>5</td>
<td>80.2</td>
<td>weathered andesite</td>
<td>70% gray</td>
<td>🔄</td>
</tr>
<tr>
<td>6</td>
<td>89.8</td>
<td>oxidized scoria</td>
<td>tuscan red</td>
<td>🔄</td>
</tr>
<tr>
<td>7</td>
<td>98.9</td>
<td>oxidized scoria</td>
<td>poinsettia red</td>
<td>🔄</td>
</tr>
<tr>
<td>8</td>
<td>108.5</td>
<td>tephra</td>
<td>fire red</td>
<td>🔄</td>
</tr>
<tr>
<td>9</td>
<td>119.0</td>
<td>tephra</td>
<td>fire red</td>
<td>🔄</td>
</tr>
<tr>
<td>10</td>
<td>131.4</td>
<td>live vegetation</td>
<td>fir green</td>
<td>🔄</td>
</tr>
<tr>
<td>11</td>
<td>145.6</td>
<td>live vegetation</td>
<td>leaf green</td>
<td>🔄</td>
</tr>
<tr>
<td>12</td>
<td>163.0</td>
<td>live vegetation</td>
<td>leaf green</td>
<td>🔄</td>
</tr>
<tr>
<td>13</td>
<td>192.7</td>
<td>dead vegetation</td>
<td>light olivenite</td>
<td>🔄</td>
</tr>
</tbody>
</table>

Application of the colors in table 5.3 to the classified raster produces the image shown in figure 6.3. Here, the colors are set to be 30% transparent, and are shown over the DEM derived hillshade image. The Mokst Butte lava flows are clearly visible, as is the
Figure 5.3: Interpreted False-Color Image of Iso Classification from Mokst Butte Study Area. See table 5.3 for color key and text for explanation.

Figure 5.4: Annotated False-Color Image of Iso Classification from Mokst Butte Study Area. See table 5.3 for color key, and text for explanation.
Figure 5.5: *Mokst Butte Cone Area False-Color Image*. Features labeled: a) scoria, black in color, in-situ; b) scoria, black in color, moved/tumbled by mass wasting; c) oxidized scoria, red in color, in-situ; d) oxidized scoria, red in color, moved/tumbled by mass wasting; g) weathered andesite, in-situ; j) conifer forest. See table 5.3 for color key and text for further explanation.

Figure 5.6: *Mokst Butte Proximal Lava Flows False-Color Image*. Features labeled: i) ash- and lapilli-sized tephra; k) shrubland; l) pine needles; h) rafted cone material; e) weathered basaltic andesite lava, in-situ; f) basaltic andesite lava, disturbed by logging operations.
surrounding vegetated area. Red areas on the lava flows are rafted cone material from syn-
eruptive collapse of the cone. Also visible in this overview of the study area is the orange 
shaded tephra produced during the Mokst Butte eruption. Figure 5.4 shows the same area, 
with these areas of interest annotated.

A closer view of the Mokst Butte cone area introduced in figure 2.4 here reveals greater 
detail in the false-color image, figure 5.5. Areas identified in field work are annotated again 
here, and show excellent agreement with the interpretations given in table 5.3. Other areas 
of interest include andesite lava areas shown in gray, and areas of shifted, sliding material on 
the oversteepened slopes of the Mokst Butte scoria cone shown in blue.

The Mokst Butte proximal lava flows introduced in figure 2.5 are reproduced here in 
the false-color image, figure 5.6. The areas identified in field work are again annotated, 
and again show excellent agreement with the lidar derived interpretations from table 5.3. 
Also notable in this image are multiple linear features colored blue, which were identified in 
field work as broken and disturbed lava resulting from equipment movements during logging 
operations. This activity (of nearly a hundred years ago) caused unweathered lava surfaces 
to be exposed to the lidar laser pulses. Though quite evident in this image, these features 
were nearly indistinguishable from the surrounding lava during field study. Also of interest 
is the thicker lava flow in the top left of the image shown mostly in gray and red colors in 
comparison with the thinner lava flow shown mostly in black in the center of the image. The 
thicker flow has a higher silica content and was emplaced later than the thinner flow, and 
in several locations overlies it. The red colors in the later flow would indicate the presence 
of oxidized scoria, and the gray color would indicate that this flow is higher in silica content 
than the earlier flow. These indications are in fact valid as the later flow is composed of 
andesite (confirmed by sample analysis) and contains abundant rafted material from the 
collapse of the Mokst Butte cone, while the earlier flow is basaltic andesite (also confirmed 
by sample analysis) and does not contain heavily oxidized material.
CHAPTER 6
DISCUSSION

6.1 Application

To illustrate an application of the process presented, the same techniques were applied to an area in the Northwest Rift Zone covering the extent of the Forest Road Flow. This is a small basalt flow emanating from fissure vents located just south of the Mokst Butte study area. Clustering and silhouette analysis indicated that 9 classes were represented in the data, and this was specified when running the Iso Cluster Unsupervised Classification tool. The initial output from this step is shown in the default colors in figure 6.1. Analysis of the lidar reflectivity values from each class are shown in table 6.1 and in the boxplot in figure 6.2.

Figure 6.1: Raw Output from Iso Cluster Unsupervised Classification in ArcGIS for the Forest Road Flow.
Table 6.1: Summary Statistics of Lidar Reflectivity by Class.

<table>
<thead>
<tr>
<th>class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>min</td>
<td>0</td>
<td>34</td>
<td>57</td>
<td>71</td>
<td>89</td>
<td>106</td>
<td>125</td>
<td>144</td>
<td>135</td>
</tr>
<tr>
<td>Q1</td>
<td>7</td>
<td>43</td>
<td>61</td>
<td>76</td>
<td>94</td>
<td>112</td>
<td>130</td>
<td>149</td>
<td>174</td>
</tr>
<tr>
<td>median</td>
<td>17</td>
<td>49</td>
<td>64</td>
<td>72</td>
<td>85</td>
<td>103</td>
<td>121</td>
<td>140</td>
<td>160</td>
</tr>
<tr>
<td>Q3</td>
<td>27</td>
<td>53</td>
<td>68</td>
<td>81</td>
<td>98</td>
<td>117</td>
<td>135</td>
<td>154</td>
<td>181</td>
</tr>
<tr>
<td>max</td>
<td>36</td>
<td>62</td>
<td>72</td>
<td>89</td>
<td>109</td>
<td>129</td>
<td>146</td>
<td>170</td>
<td>254</td>
</tr>
<tr>
<td>mean</td>
<td>17.0</td>
<td>47.9</td>
<td>64.3</td>
<td>80.5</td>
<td>98.4</td>
<td>116.7</td>
<td>134.9</td>
<td>155.1</td>
<td>185.1</td>
</tr>
<tr>
<td>st dev</td>
<td>10.68</td>
<td>6.27</td>
<td>4.30</td>
<td>5.07</td>
<td>5.25</td>
<td>5.21</td>
<td>5.60</td>
<td>6.67</td>
<td>14.59</td>
</tr>
</tbody>
</table>

Figure 6.2: Boxplot of Reflectivity Values by Class for Forest Road Flow. Note: statistics used for this plot and which are shown in Table 6.1 above are reflectivity values only, while classification is based on curvature and height data as well as reflectivity.

Colors for the classes were chosen using the same criteria used for the Mokst Butte study area (see Table 5.1). The classes, interpretations, and assigned colors are shown in Table 6.2 below, and again, are based solely on reflectivity values, while the classification is based on a combination of reflectivity, curvature, and height data. Here, the reduced number of appropriate classes indicated by the silhouette analysis is not unexpected in that there is no scoria cone or associated material at this lava flow, nor is there any andesite lava. This situation makes clear the need for a quantitative determination of the number of classes present in the data. It should be noted that class 4 has a mean reflectivity of 80.5, which would indicate that it should be categorized as andesite lava. However, when colored that way, it was clear...
that this class was simply a small band of noise in the data, wholly surrounded by the basalt of the lava flow. As a result, this class was colored black to match the surrounding material.

Table 6.2: Mean Reflectivity, Inferred Terrain Type, and Color Assignment by Class.

<table>
<thead>
<tr>
<th>Class</th>
<th>Mean Reflectivity</th>
<th>Inferred Terrain Type</th>
<th>ArcGIS Color</th>
<th>Color Swatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17.0</td>
<td>specular surfaces</td>
<td>yogo blue</td>
<td>![Swatch]</td>
</tr>
<tr>
<td>2</td>
<td>47.9</td>
<td>glassy/fresh basalt</td>
<td>ultra blue</td>
<td>![Swatch]</td>
</tr>
<tr>
<td>3</td>
<td>64.3</td>
<td>weathered basalt</td>
<td>black</td>
<td>![Swatch]</td>
</tr>
<tr>
<td>4</td>
<td>80.5</td>
<td>weathered basalt</td>
<td>black</td>
<td>![Swatch]</td>
</tr>
<tr>
<td>5</td>
<td>98.4</td>
<td>tephra</td>
<td>fire red</td>
<td>![Swatch]</td>
</tr>
<tr>
<td>6</td>
<td>116.7</td>
<td>tephra</td>
<td>fire red</td>
<td>![Swatch]</td>
</tr>
<tr>
<td>7</td>
<td>134.9</td>
<td>living vegetation</td>
<td>fir green</td>
<td>![Swatch]</td>
</tr>
<tr>
<td>8</td>
<td>155.1</td>
<td>living vegetation</td>
<td>leaf green</td>
<td>![Swatch]</td>
</tr>
<tr>
<td>9</td>
<td>185.1</td>
<td>dead vegetation</td>
<td>light olivenite</td>
<td>![Swatch]</td>
</tr>
</tbody>
</table>

Figure 6.3: Interpreted False-Color Image of Iso Classification of Forest Road Flow. See table 6.2 for color key and text for explanation.
6.2 Conclusions

Limitations of the “clustering/silhouette analysis/self-organizing map” method include potential failure to properly classify features when extrapolation is attempted beyond the data used in the clustering and silhouette analysis. And although simple random sampling is effective in allowing much larger study areas to be evaluated, difficulties may still arise due to the size of the datasets and the computationally expensive algorithms being used.

In spite of these relatively minor drawbacks, the methods and process described allow the maximum amount of information to be extracted from the type lidar data that is becoming more and more readily available. Not only are basic volcanic terrains and morphologies such as lava flows and vegetated areas identified, but subdivisions within those terrains are identified, some of which are not identifiable in other commonly used remote sensing data such as color orthophotography and InSAR (synthetic aperture radar). Differences in lava composition, the presence of rafted scoria, tephra deposits in vegetated areas, and lava which has been moved since primary emplacement are all clearly and accurately identified. Being able to discern and identify terrains and morphologies remotely has obvious utility in mapping and the study of remote or inaccessible regions. As a tool to direct field work to areas of interest, the savings in time, expense, and exposure to personal risk could be substantial.
REFERENCES


Script to perform initial processing of data

dataframe = '8';

% load data and construct data matrix with one row per cell
data = [reshape(dlmread(['df' dataframe '_data_ref.txt'])), [], 1],...
reshape(dlmread(['df' dataframe '_data_crv.txt'])), [], 1),...
reshape(dlmread(['df' dataframe '_data_hts.txt'])), [], 1]);

% remove rows containing nulls from data
% find and remove rows with reflectivity nulls
cond = (data(:,1) == -9999);
data(cond,:) = [];
% find and remove rows with curvature nulls
cond = (data(:,2) == -9999);
data(cond,:) = [];
% find and remove rows with height nulls
cond = (data(:,3) == -9999);
data(cond,:) = [];
clear cond;

% normalize reflectivity data to 0-254 range
(data(:,1)-min(data(:,1))) * (254/(range(data(:,1))));
% process total curvature data
% shift all values so min is equal to smallest positive value
cond = data(:,2) > 0;
data(:,2) = data(:,2) + min(data(cond,2));
% take log of data values
data(:,2) = log(data(:,2));
% shift all values so min is zero
data(:,2) = data(:,2) - min(data(:,2));

% process height data
% shift all values so min is equal to smallest positive value
cond = data(:,3) > 0;
data(:,3) = data(:,3) + min(data(cond,3));
% take log of data values
data(:,3) = log(data(:,3));
% shift all values so min is zero
data(:,3) = data(:,3) - min(data(:,3));

% save transformed data matrix to file
save(['df' dataframe '_final_data.txt', 'data'])

% plot reflectivity data
figure;
hist(data(:,1),(range(data(:,1))+1));
title(['df' dataframe ' final lidar reflectivity data']);
xlim([min(data(:,1)) max(data(:,1))]);
xlabel('transformed reflectivity value');
ylabel('cell count');

% plot transformed curvature data
figure;
hist(data(:,2),254);
title(['df' dataframe ' final total curvature data']);
xlim([min(data(:,2)) max(data(:,2))]);
xlabel('transformed log curvature value');
ylabel('cell count');

% plot transformed height data
figure;
hist(data(:,3),254);
title(['df' dataframe ' final height data']);
xlim([min(data(:,3)) max(data(:,3))]);
xlabel('transformed log height (feet)');
ylabel('cell count');

% show table of summary statistics
colnames = {'reflectivity','curvature','height'};
T = summary(data(:,:),colnames);
% convert table to string form
TString = evalc('disp(T)');
% use TeX Markup for bold formatting and underscores
TString = strrep(TString,'<strong>','\bf');
TString = strrep(TString,'</strong>','\rm');
TString = strrep(TString,'_','\_');
% get a fixed-width font
fixedWidth = get(0,'FixedWidthFontName');
% output the table using the annotation command
figure;
annotation(gcf,'Textbox','String',TString,'Interpreter','Tex',...
'FontName',fixedWidth,'Units','transformed','Position',[0 0 1 1]);

Script to generate phylogenetic trees and perform silhouette analysis

data = importdata('df8_final_data.txtdata');
dataframe = '8';
numSamples = 5000; reps = 5;
minK = 2; maxK = 20;
random = [5 6 12 23];
for p = 1:4
    rngID = num2str(random(p));
silEval = zeros(reps,(maxK-minK+1));
    rng(str2double(rngID));
    idx = randperm(length(data.data));

    for j = 1:reps
        phyFile = ['df' dataframe '_experiment' num2str(p) '_rng' rngID '_rep'...
        num2str(j) '_phyplot.png'];
        phydata = data.data(idx((j*numSamples+1)-numSamples:j*numSamples),:);
        clusteval = evalclusters(phydata, 'linkage', 'silhouette',...
        'distance', 'euclidean', 'ClusterPriors', 'equal', 'KList', minK:maxK);
        evalplot = clusteval.CriterionValues;
silEval(j,:) = evalplot;

Z = linkage(phydata, 'ward', 'euclidean');
phy = phytree(Z);
h = plot(phy, 'type', 'EqualAngle', 'LeafLabels', false,...
'TerminalLabels', false);
set(h.axes, 'xtick', [], 'xticklabel', '', 'ytick', []);
set(h.BranchDots, 'marker', 'none');
set(h.LeafDots, 'marker', 'o', 'markersize', 4, 'markerfacecolor', 'k');
title(['experiment ' num2str(p) '; n = ' num2str(numSamples)...
'; rep # ' num2str(j)])];
saveas(gcf,phyFile);
while ( ~exist(phyFile,'file') )
sleep(1);
end
delete(gcf);
phyDataFile = ['df' dataframe '_experiment' num2str(p) '_rng' rngID...
'_rep' num2str(j) '_phydata'];
save(phyDataFile, 'phy');
end
silEvalFile = ['df' dataframe '_experiment' num2str(p) '_rng' rngID '_sildata'];
save(silEvalFile, 'silEval')
meanEval = mean(silEval);

hold on
plot(silEval,'k--')
plot(meanEval,'k','LineWidth',3)
xlin([1 (maxK-minK+1)]); xticks(1:(maxK-minK+1));
ylin([.35 .60]);
xticklabels(minK:maxK);
title(['experiment ' num2str(p) '; n = ' num2str(numSamples)...'
'; ' num2str(reps) ' reps']);
xlabel('number of clusters'), ylabel('silhouette value');
plotFile = ['df' dataframe '_experiment' num2str(p) '_rng' rngID '_silplot.png'];
saveas(gcf,plotFile);
while (~exist(plotFile,'file'))
sleep(1);
end
delete(gcf);
end

Script to perform analysis of reflectivity values by class

classes = 13; digits = 2;
colnames = cell(1,classes); stats = table();

% load data and compute summary statistics for each class
for i = 1:classes
    colnames{i} = {{'class' num2str(i)}];
data = reshape(dlmread...
    ([{'ref_class' num2str(i) '.txt'}], [], 1);
data(data == -9999) = [];
dataStats = sumStats(data,colnames{i},digits);
stats = [stats dataStats];
end
clear data;

% save summary stats to ascii file
statsFile = stats{:,:};
save('ref_stats.txt','statsFile','-ascii');

% display table of summary statistics by class
% convert table to string form
TString = evalc('disp(stats)');
% use TeX Markup for bold formatting and underscores
TString = strrep(TString,'<strong>','\bf');
TString = strrep(TString,'</strong>','\rm');
TString = strrep(TString,'_','\_');
% get a fixed-width font
fixedWidth = get(0,'FixedWidthFontName');
% output the table using the annotation command
figure;
annotation(gcf,'Textbox','String',TString,'Interpreter','Tex',... 'FontName',fixedWidth,'Units','Normalized','Position',[0 0 1 1]);

% create side-by-side boxplot of summary statistics
plotstats = stats{1:5,:};
figure;
boxplot(plotstats,'widths',0.7)
title('boxplot of reflectivity statistics by class');
xlabel('class'); ylabel('reflectivity value');

Function to compute summary statistics
function [table] = sumStats(x,colnames,digits)
if nargin < 2, error('not enough input arguments'); end
if nargin == 2, digits = 1; end

%%% this function returns a table of summary statistics
% x data may be a vector in row or column form,
% or a matrix, but in column form only
% colnames must be a cell array of column names
% digits is the number of decimal places (default is 1)
stats = [min(x); prctile(x,25); median(x);...
prctile(x,75); max(x); mean(x); std(x)];
stats = round(stats,digits);
rownames = {'min';'Q1';'median';'Q3';'max';'mean';'st dev'};
table = array2table(stats,'rownames',rownames,'variablenames',colnames);
end