The University of Pittsburgh

Spatial Cluster Analysis for landslide Activity in Western Pennsylvania

Tom Plazek

Table of Contents

Spatial Cluster Analysis for landslide Activity in Western Pennsylvania 0
List of Figures 1
Abstract 2
Introduction 2
Methods 2
3.1 Georeferencing 2
3.2 Digitizing 2
3.3 Grouping and Zonal Statistics 2
3.4 Cluster Analysis 3
3.5 Joining Cluster Data 3
Results and Discussion 3
Table 1 – Summary of cluster counts across the landscape 5
Conclusion 7
Appendix A – Tables 8
Appendix 1a – Slide types 8
Appendix 2a- Annotation codes 8
Appendix 3a – Landslide type to count data (summarize) 9
Appendix B – R source code 9
Appendix1b – R cluster source code 10
Appendix 2b – Fig. 2 source code (R) 11
Appendix C – Supplementary maps 12

List of Figures

Figure 1 - Dendrogram of the hierarchical clustering in R 4
Figure 2 - An R-plot visualizing the landslide count to landslide-type data 5
Figure 3 - Regional landslide map classified by landslide cluster values 6
Figure 4 - Counts by cluster pie chart. This chart indicates that 99.6% of all slides were categorized as being in a region with only one cluster. 7
Figure 5 - Appendix 1c: Coalport, PA layout with classified slide types 12
Figure 6 - Appendix 2c: Group data layout without graticule 13

Abstract
Using digitized Quadrangle maps produced by the US Geologic Survey which indicated locations and types of landslides across the state, a cluster analysis was run to determine groupings of landslides. Zonal statistics were calculated in ArcGIS for different landscape features. Using R, hierarchical clustering was run on these normalized statistical outputs to determine if there was any spatial clustering concerning landslides. Only about nine percent of the landslides were determined to have their origin in manmade features, indicating mostly natural geomorphological processes working. Additionally, most of the landslide features were spatially grouped in an area that was determined to only contain one cluster. Over ninety-nine of all the digitized landslides were grouped this way. This was an indicator that while the mobile regolith is active in this region, the magnitude of these events may be generally small.

Introduction
Western Pennsylvania is a region that is known to be prone to landslides. Quadrangle maps from the area, produced by the United States Department of the Interior Geological Survey in the 1950s are relatively extensive maps that outline areas known to have or are susceptible to landslides. These maps include the type of slide, along with a specific corresponding graphic. Slides in manmade features (strip mines, coal refuse banks, quarries, gravel pits) are accompanied by an annotation that indicates characteristics at each site. The quadrangles were digitized in ArcGIS, peer-reviewed, and combined into one landscape. The curvature, slope, geomorphon, Topographic Moisture Index (TMI), and Terrain Ruggedness (TRI) were calculated, and zonal statistics were run on them. These statistics allowed for cluster analysis to be executed in R. *It is hypothesized that most of these slides were caused by human activity and that density-based clustering techniques will show this.*

Methods

3.1 Georeferencing

For this work, Quadrangle maps created by the United States Department of the Interior Geological Survey were used. The PDF version of the map first was converted to a TIFF file, to be compatible as a raster format in ArcGIS Pro. Since the new TIFF files lacked a coordinate reference system (CRS), they needed to be georeferenced. Using the 7.5-minute series graticules, along with recognizable geographic and/or geologic features, control points were established. From the control points, the TIFF could be overlain in the appropriate area.

3.2 Digitizing

Once the CRS had been rectified, the existing features shown on the map were digitized. Two distinct feature classes for lines and polygons were created. A numeric indicator that corresponded to the slide type was created in the attribute table, along with two Excel tables containing the slide type information and annotation codes, respectively. These tables were then joined to the completed line and polygon feature classes (Appendix 1a and 1b).

3.3 Grouping and Zonal Statistics

Once the digitization was complete, each person exported their line and polygon shapefiles, which were combined into a large region of Western Pennsylvania. The curvature, slope, geomorphon, TMI, and TRI were calculated, as well as the tabular area. Curvature values

indicate which parts of the surface are concave or convex, and to what degree. Slope values indicate the angle of the hill. Geomorphon is a digital terrain algorithm that maps and classifies landforms based on their morphology. Using shape and slope, it can class a region into nine landform types: Flat, Peak, Ridge, Shoulder, Slope, Hollow, Footslope, Valley, and Pit. TMI values are dimensionless numbers that estimate the water balance at each location by comparing upslope contributing areas to the topographic wetness index. TRI is a terrain ruggedness index that expresses the surface complexity in meters. Tabular area calculates the area of each polygon that intersects with each zone in a raster. A table is then generated which summarizes the areas of each polygon by zone. Summary statistics, such as mean, maximum, minimum, and standard deviation are also calculated and included in the table. Minima and maxima for all layers were calculated using the Zonal Statistic tool in ArcGIS Pro. The output tables were then exported as CSV files.

3.4 Cluster Analysis

The CSVs containing the zonal statistic outputs were loaded into a prefabricated R script. This source code can be found in the appendix (Appendix 3). The values were bound into a sorted matrix and hierarchical clustering was run. A dendrogram was then created and a CSV file was exported with the new cluster data (fig. 1).

3.5 Joining Cluster Data

The exported cluster data was joined to the class shapefile in ArcGIS and visualized using unique values. The cluster join was then summarized by count (table 1). A bar plot was created to visualize this summarize (fig. 2).

Results and Discussion

Georeferencing proved to be difficult for many people, not only in matching up control points to real-world features but also in being confused by the different polynomial stretches. Once this was sorted out, the shapefiles were able to be exported and combined. Not everyone's data was included in the dataset, due to the sheer size of the project. Dr. Bain combined the shapefiles, as well as calculated the curvature, slope, geomorphon, TMI, and TRI. Due to download issues with OneDrive, many students were not able to run zonal statistics on their own. Robert Murphy allowed the class to use his statistical outputs. This also kept the statistics very consistent since the same datasets were being used by the entire class.

The R-program which was used required the 'cluster' and 'dplyr' packages. All of the zonal statistics tables needed to be joined together with a left join (via ID) to create a combined table. Null values were removed on this combined table and it was scaled into a matrix. Using 'cbind', columns were combined and standardized using the 'scale' function. This created a new data frame that subtracted the mean of each column from each value in the column and then divided it by the standard deviation of that same column. The standardized data frame was then combined with the original using 'cbind'. This was an important step to complete before running the hierarchical clustering. Clustering requires data to be standardized to ensure that different variables are measured on comparable scales.

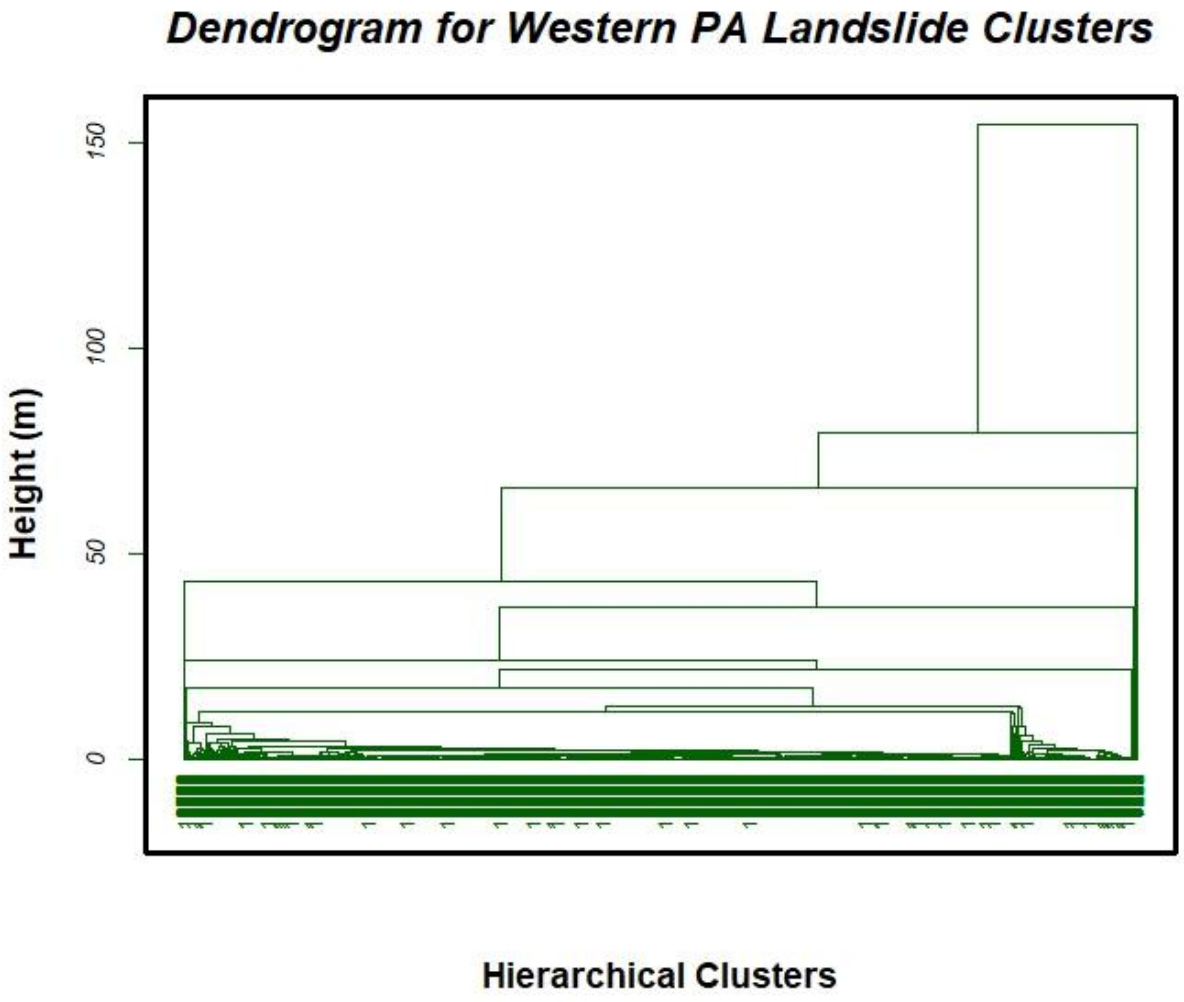

Figure 1 - Dendrogram of the hierarchical clustering in R

OBJECTID	Clusters below 50m	FREQUENCY	Cluster Count
1	--	1040	--
2	1	8983	8983
3	2	17	17
4	3	4	4
5	4	2	2
6	5	1	1
7	6	3	3
8	7	3	3
9	8	2	2

Table 1 – Summary of cluster counts across the landscape

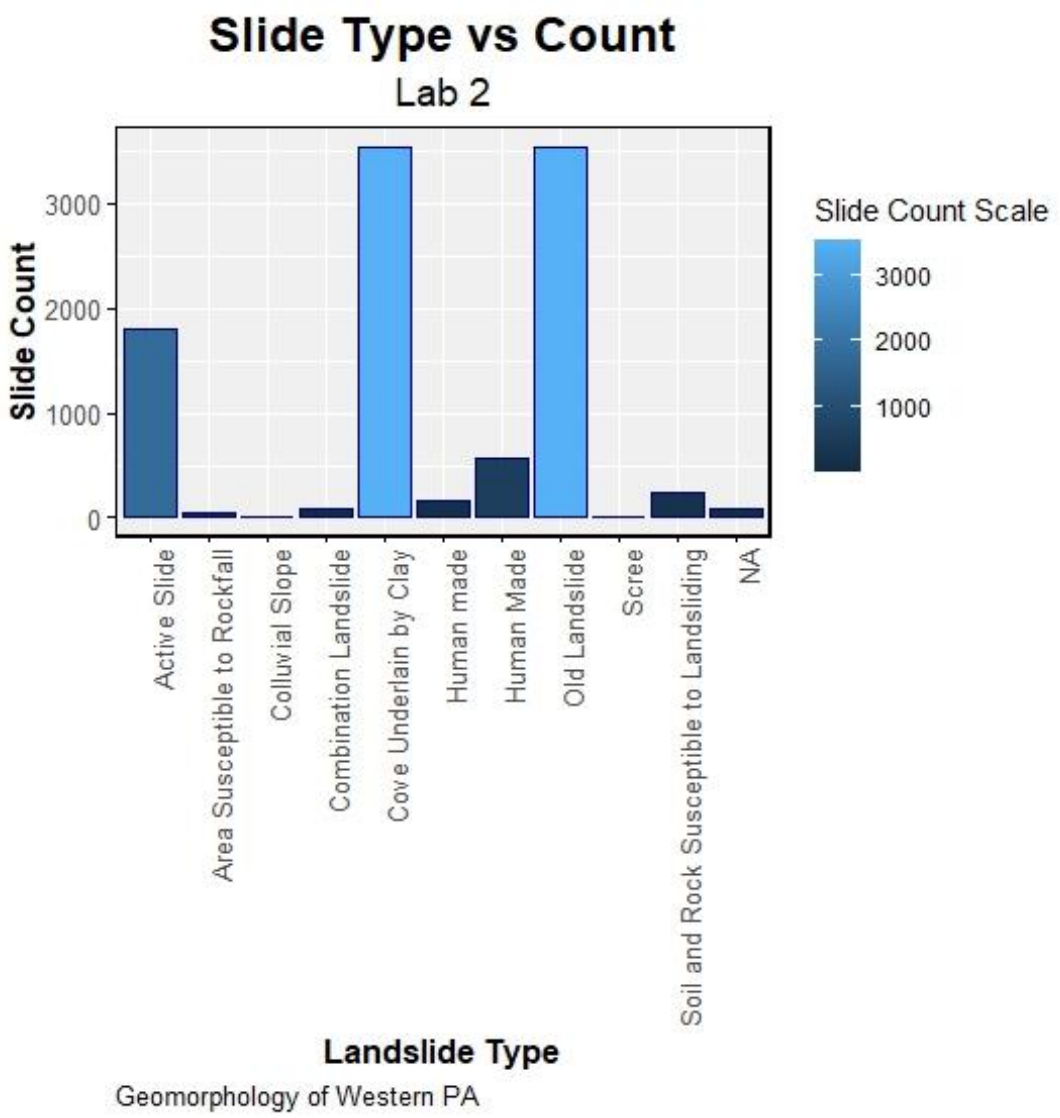

Figure 2 - An R-plot visualizing the landslide count to landslide-type data

R uses agglomerative clustering, as opposed to divisive clustering. The former method starts with each data point as it one cluster and then merges them in iterations to the closest clusters It does this until every data point is assigned to a single cluster. Euclidean distance was used as the distance metric to cluster variables contained in the newly standardized matrix. Using the cluster outputs from 'hclust', a dendrogram was created which visualized the number of clusters relative to landscape height (see fig. 1). The 'cutree' function was used to extract only the clusters under 70 meters. These clusters were then exported to ArcGIS and joined to the class data.

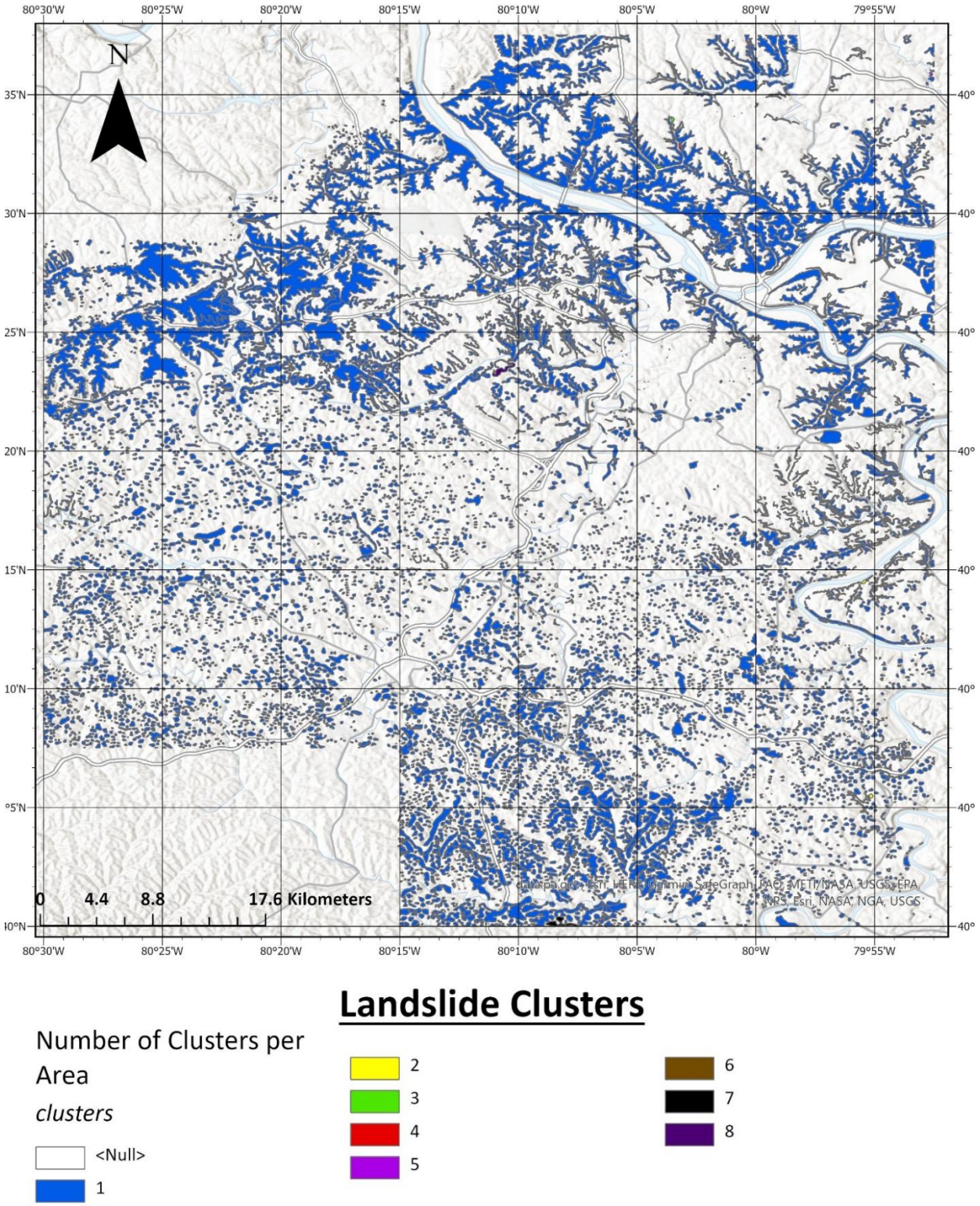

Figure 3 - Regional landslide map classified by landslide cluster values

The exported cluster outputs displayed a varied landscape in ArcGIS. Eight clusters were extracted under a height of 70 meters. The vast majority of the clusters fell into the '1' category, indicating one cluster per area (see Fig. 2). The classified cluster map can be viewed below (Fig.3). A count vs slide type plot was also created to visualize the number of each landslide type in the region (see Fig. 2). 'Old landsides' and 'Cover underlain by clay' were by far the most prevalent types of slides each having over 3,500 occurrences. The second most counted type was 'Active landslides' which has a count of 1,794. Human-made features made up only about nine percent of the total landslide count. This indicates that the vast majority of the landslides occurred naturally, making the mobile regolith in this region very active. Natural landslide counts were an order of magnitude larger than the manmade features. All cluster regions must made up mostly of natural features. Most of the regions only contained one cluster, however; this means that, while active, the magnitude of the geomorphic processes may be relatively small. A larger magnitude should relate to a greater number of clusters in the region. Over 99% of the region was clustered in a '1' grouping. This shows relatively low-magnitude incidents (Fig. 4).

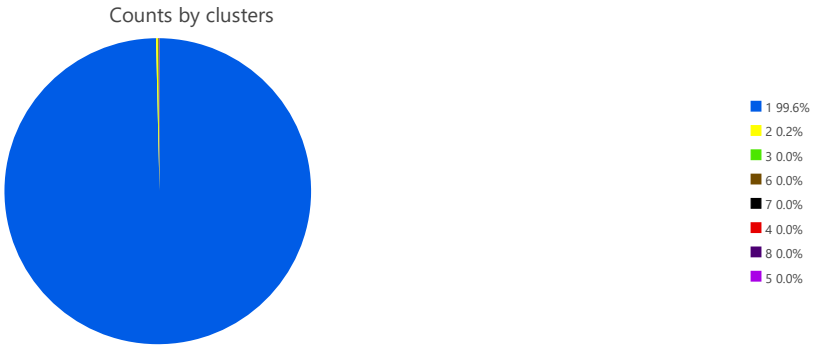

Figure 4 - Counts by cluster pie chart. This chart indicates that 99.6% of all slides were categorized as being in a region with only one cluster.

Conclusion

Overall, it was determined that the majority of landslide activity originates from natural processes. Human activity only created around nine percent of all landslides in the region. While these landslide occurrences are very common in Western PA, the magnitude seems to be small. Over 99% of the region's areas were classified in a '1'cluster group, indicating only one landslide cluster per given area.

Special thanks to Robert Murphy for running the zonal statistics and sharing the results with the class

Appendix A – Tables

ID	TYPE
1	Active/recently active landslides
2	Old Landslides
3	Combination Landslides
4	Colluvial Landslides
5	Scree
6	Colluvial slopes with landslides
7	Areas susceptible to debris flow/avalanches
8	Areas susceptible to rock fall (line data)
9	Soil and rock susceptible to land sliding
10	Strip mines
11	Coal refuse banks
12	Quarries
13	Gravel pits
14	Slides in human-made features

Appendix 1a – Slide types

Annotation_code	Annotation
SH	Bench with high wall: strip mine
SF	Furrowed with high wall: strip mine
SD	Multiple furrows and multiple benches: strip mine
SS	Hilltop removed: strip mine
SRG	Reclaimed by grading: strip mine
SRU	Reclaimed by secondary use: strip mine
SH/R	Regraded in part; high wall remains: strip mine
R	Coal refuse bank identified on areal photo, not classified in field check
RB	Coal refuse bank not burnt, nor on fire
RBB	Coal refuse bank burnt
RBD	Coal refuse bank burning
RBS	Coal refuse bank sludge
Q	Quarry site
QUB	Spoil bank, quarry waste
G	Site of gravel pit
AF	Earth flow in fill : man-made feature
A/S	Earth flow in strip castings: man-made feature
A/R	Earth flow in coal refuse: man-made feature

Appendix 2a- Annotation codes

OBJECTID	Landslide Type	FREQUENCY	COUNT
1	-------	83	83
2	Active Slide	1794	1794
3	Area Susceptible to Rockfall	51	51
4	Colluvial Slope	10	10
5	Combination Landslide	80	80
6	Cove Underlain by Clay	3541	3541
7	Human-made	165	165
8	Human Made	562	562
9	Old Landslide	3530	3530
10	Scree	7	7
11	Soil and Rock Susceptible to Landsliding	232	232

Appendix 3a – Landslide type to count data (summarize)

Appendix B – R source code

#This script uses hierarchical clustering to cluster Western PA Landslide data and creates a dendrogram to visualize the results.

```
setwd("C:/Users/Thoma/OneDrive - University of Pittsburgh/Spring Semester 2023/GEOL 1060 Geomorphology/Lab/Lab 2/Group Data")
# Load packages
library(cluster)
library(dplyr)
library(rmarkdown)
library(knitr)
library(sf)
# Read data
MinMaxCurvatureCSV <- ("C:/Users/Thoma/OneDrive - University of Pittsburgh/Spring Semester 2023/GEOL 1060 Geomorphology/Lab/Lab 2/Group Data/OneDrive_1_4-3-2023/CSVs/MinMaxCurvatureCSV.csv")
MinMaxSlopeCSV <-read.csv("C:/Users/Thoma/OneDrive - University of Pittsburgh/Spring Semester 2023/GEOL 1060 Geomorphology/Lab/Lab 2/Group Data/OneDrive_1_4-3-2023/CSVs/MinMaxSlopeCSV.csv")
MinMaxTMICSV <- read.csv("C:/Users/Thoma/OneDrive - University of Pittsburgh/Spring Semester 2023/GEOL 1060 Geomorphology/Lab/Lab 2/Group Data/OneDrive_1_4-3-2023/CSVs/MinMaxTMICSV.csv")
MinMaxTRICSV <- read.csv("C:/Users/Thoma/OneDrive - University of Pittsburgh/Spring Semester 2023/GEOL 1060 Geomorphology/Lab/Lab 2/Group Data/OneDrive_1_4-3-2023/CSVs/MinMaxTMICSV.csv")
TabAreaLanslideCSV <- read.csv("C:/Users/Thoma/OneDrive - University of Pittsburgh/Spring Semester 2023/GEOL 1060 Geomorphology/Lab/Lab 2/Group Data/OneDrive_1_4-3-2023/CSVs/TabAreaLanslideCSV.csv")

#check CRS
st_crs(MinMaxCurvatureCSV)
```

```
st_crs(MinMaxSlopeCSV)
st_crs(MinMaxTMICSV)
st_crs(MinMaxTRICSV)
St_crs(TabAreaLanslideCSV)
# Join data
zstat_1 <- left_join(TabAreaLanslideCSV, MinMaxCurvatureCSV, by = c("ID" = "Id"));
zstat_2 <- left_join(zstat_1, MinMaxSlopeCSV, by = c("ID" = "Id"));
zstat_3 <- left_join(zstat_2, MinMaxTMICSV, by = c("ID" = "Id"));
clusterdata <- left_join(zstat_3,MinMaxTRICSV, by = c("ID" = "Id"))
#Remove NA/no data
clusterdata<-na.omit(clusterdata)
#cbind into scaled matrix
clusterdata2<-cbind(scale(cbind(clusterdata[,3:12],clusterdata[,16:18])))
#compute distance matrix
d <- dist(clusterdata2)
#hierarchical cluster
hc <- hclust(d, method = "complete")
#plot dendrogram
plot(hc, main = "Dendrogram for Western PA Landslide Data",
    col = "dark blue",
    hang = -1,
    labels = clusterdata$ID,
    xlab = "Hierarchical Clustering",
    font.lab = 2)
#Extract clusters at height of 2000
clusters<-cutree(hc,h=50)
#Combine ID column and extracted clusters for each ID via cutree
polygonCluster<-cbind(TabAreaLanslideCSV[1:(nrow(TabAreaLanslideCSV)-3),1], clusters)
write.csv(polygonCluster, file="C:/Users/Thoma/OneDrive - University of Pittsburgh/Spring Semester
2023/GEOL 1060 Geomorphology/Lab/Lab 2/clusters50.csv")
```

Appendix1b – R cluster source code

```
library(ggplot2)
library(dplyr)
library(viridis)
library(RColorBrewer)

ggplot(data = landslide_type_summarize, aes(x = LandslideDataForClass_Type, y =
COUNT_LandslideDataForClass_Type,
                                fill = COUNT_LandslideDataForClass_Type)) +
  geom_bar(stat = "identity", color = "dark blue") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(
    title = "Slide Type vs Count",
    subtitle = "Lab 2",
    caption = "Geomorphology of Western PA",
    x = "Landslide Type",
```

```
    y = "Slide Count",
    fill = "Slide Count Scale") +
  theme(panel.border = element_rect(color = "Black" , fill = NA, size = 1),
        panel.background = element_rect(fill = "#f0f0f0"),
        axis.line = element_line(color = "black"),
        axis.ticks = element_line(color = "black"),
        axis.title = element_text(face = "bold", size = 12),
        axis.text = element_text(size = 10),
        plot.title = element_text(face = "bold" , size = 18, hjust = 0.5),
        plot.subtitle = element_text(size = 14, hjust = 0.5),
        plot.caption = element_text(size = 10, hjust = 0))

library(plotly)

# create ggplot object
p <- ggplot(data = landslide_type_summarize, aes(x = LandslideDataForClass_Type, y =
COUNT_LandslideDataForClass_Type,
                                                 fill = COUNT_LandslideDataForClass_Type)) +
  geom_bar(stat = "identity", color = "dark blue") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(
    title = "Slide Type vs Count",
    subtitle = "Lab 2",
    caption = "Geomorphology of Western PA",
    x = "Landslide Type",
    y = "Slide Count",
    fill = "Count") +
  theme(panel.border = element_rect(color = "Black" , fill = NA, size = 1),
        panel.background = element_rect(fill = "#f0f0f0"),
        axis.line = element_line(color = "black"),
        axis.ticks = element_line(color = "black"),
        axis.title = element_text(face = "bold", size = 12),
        axis.text = element_text(size = 10),
        plot.title = element_text(face = "bold" , size = 18, hjust = 0.5),
        plot.subtitle = element_text(size = 14, hjust = 0.5),
        plot.caption = element_text(size = 10, hjust = 0))

# convert ggplot object to plotly object
ggplotly(p)
```

Appendix 2b – Fig. 2 source code (R)

Appendix C – Supplementary maps

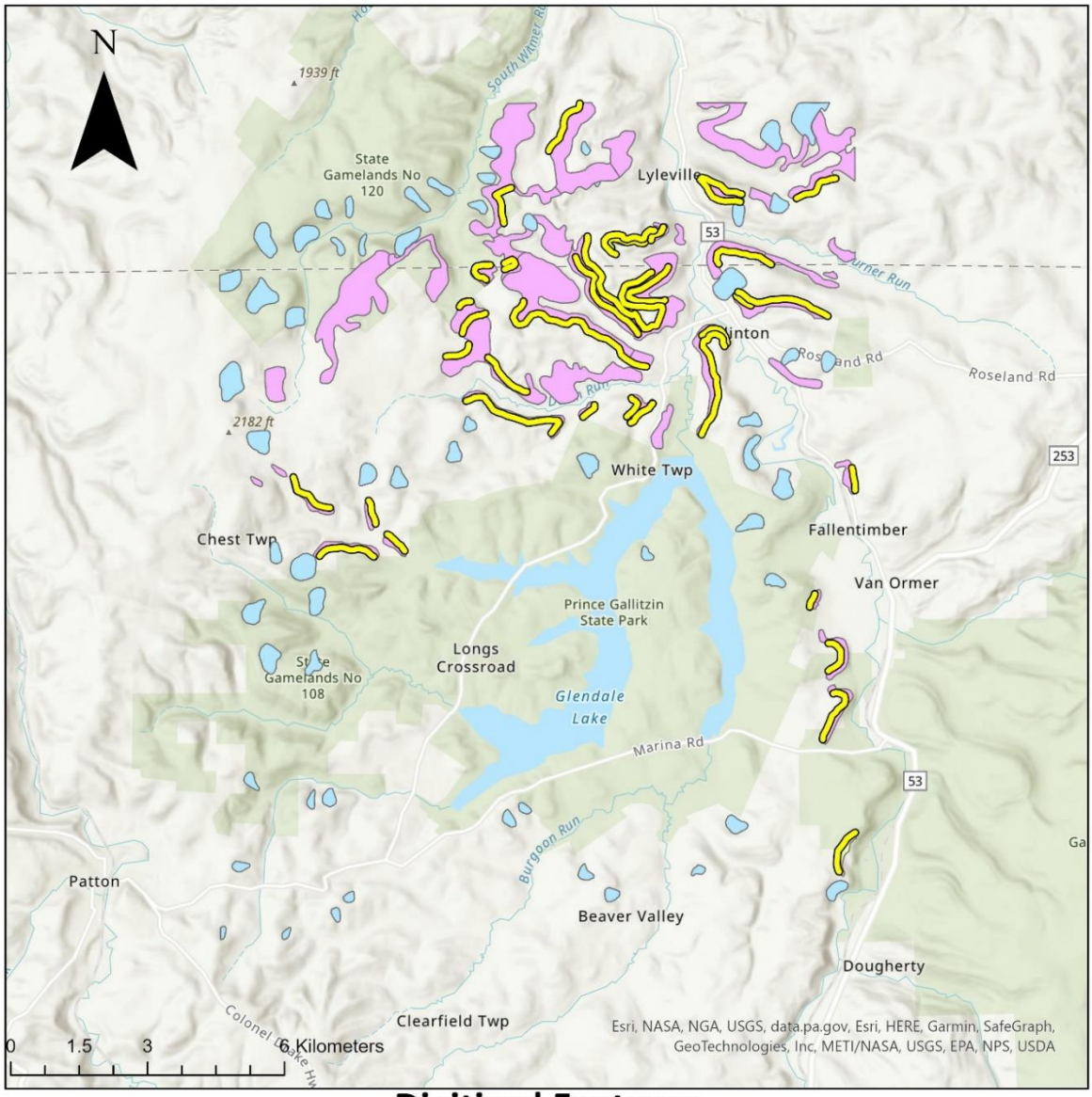

Figure 5 - Appendix 1c: Coalport, PA layout with classified slide types

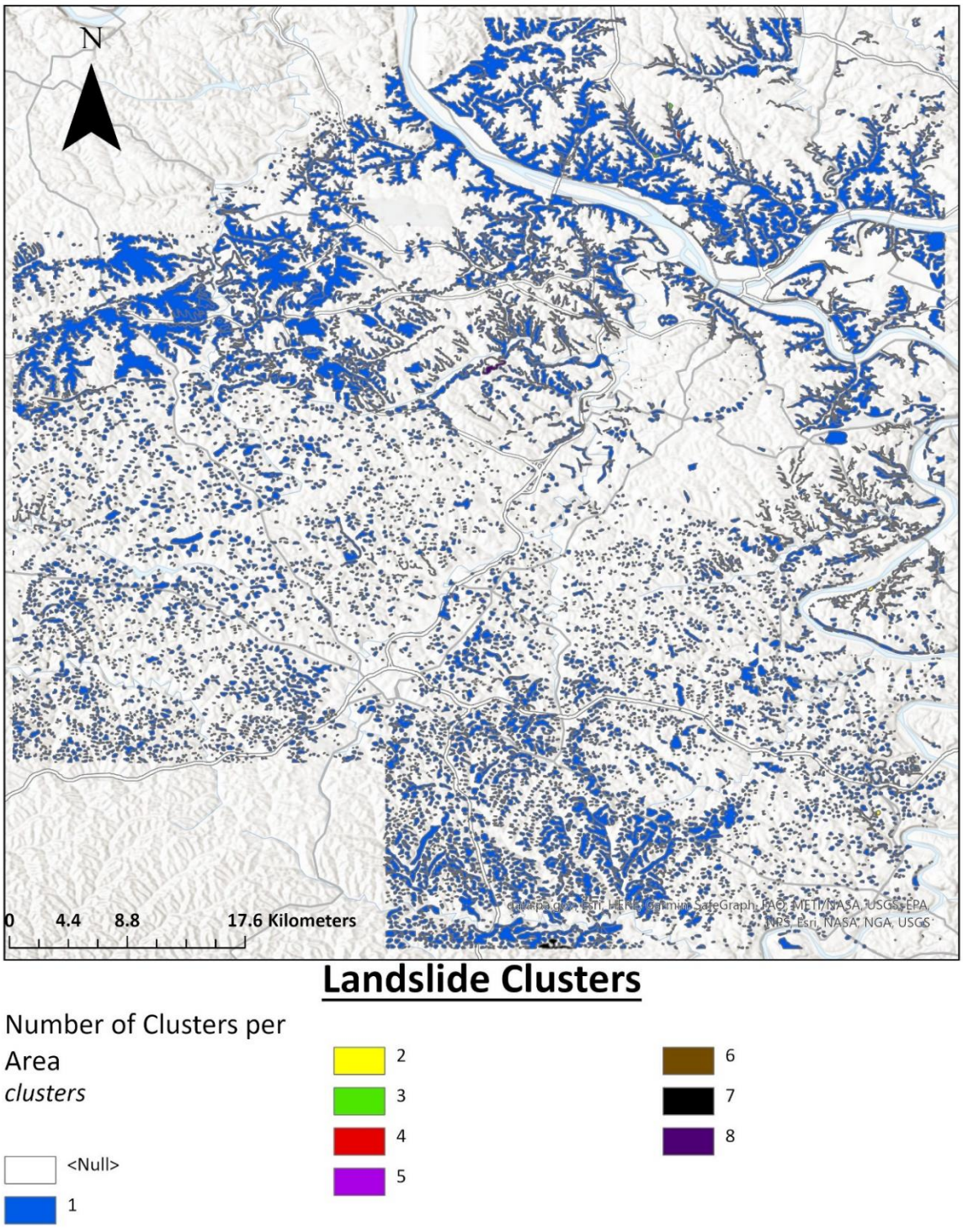

Figure 6 - Appendix 2c: Group data layout without graticule