Rio Santa Geodatabase Project

Amanda Cuellar

December 7, 2012

Introduction

The McKinney research group (of which I am a part) collaborates with international and onsite researchers to evaluate the risks posed by high mountain glacial lakes in the Rio Santa Basin in Peru. In particular, the group has taken field data for Lake Palcacocha, the Pastaruri Glacier, and the Arteson Glacier. Of particular concern for the researchers is the potential for an outburst flood from Lake Palcacocha. As temperatures rise due to global climate change, melting glaciers lead to growth of the lake and weakening of the terminal moraine (the dam that holds the lake in), which is partially made of ice. The McKinney group and other researchers have studied the glaciers, lake, and moraine in order to better understand the risk of an outburst flood, model how the flood would affect communities downstream, and evaluate options that would mitigate downstream damage or avoid a flood altogether. The collected and model generated data from the group's and collaborator's studies are stored locally, which makes sharing of the information difficult. Therefore, for this project my objective is to create a GIS map database for the data gathered by the McKinney group and collaborators for the Rio Santa Basin (which contains the glaciers under study). In addition to collecting, organizing, and standardizing data from the McKinney group and collaborators, I also created a new layer for the database showing ground cover in the Rio Santa basin.

The map database produced in this project is meant to be shared with researchers around the world and to serve as a platform for researchers to upload and share data they have collected. For this reason the geodatabase created must have a logical, easy to understand structure, all datasets must be available for download by users, and users must be able to upload their own datasets. To this end I identified five tasks for this project:

1. Gather data from collaborating researchers and the McKinney group

2. Design an organization structure for the geodatabase that facilitates access by researchers looking for data and those uploading data sets

3. Standardize coordinate systems for all datasets and integrate into one geodatabase

4. Create new layers to fill in missing information or from tabular data not in GIS format

5. Share the geodatabase with collaborators and the community of practice

This report will be organized by addressing each of the five tasks in turn. Although I was able to address all five tasks in my project some remain incomplete. I will continue this project after the class as part of my research and intend to address the shortcomings described here.

Task 1: Collecting data

For the initial version of the Rio Santa geodatabase I (in conjunction with two members from the McKinney group) reached out to current and past McKinney group members. In addition I worked with collaborators at The Mountain Institute in Peru that have been working in the Rio Santa basin for four years. I was able to communicate with the Peruvian team through Skype. The Peruvian team was able to share the datasets they had via e-mail and through dropbox.

Most of the information contributed by the Peruvian team was in the form of GIS vector and raster files. The results of surveys conducted in the area, though, were not in GIS format. These surveys were intended to understand the perception of risk, demographics, and concerns of inhabitants of the Rio Santa Basin. Therefore survey results we received were in tabular format (Excel spreadsheet) or written as individual statements in reports.

Task 2: Design an Organization Structure

I worked with members of the McKinney team to determine what data sets are important for the study of glacier lake outburst floods and to review the data sets we had received from the Peru team. Together we defined an organization system for the geodatabase that was logical and easy to understand. We then presented this organization structure to the Peruvian team to obtain their feedback as future users of the geodatabase. Through this process we created the organization structure shown in Figure 1.

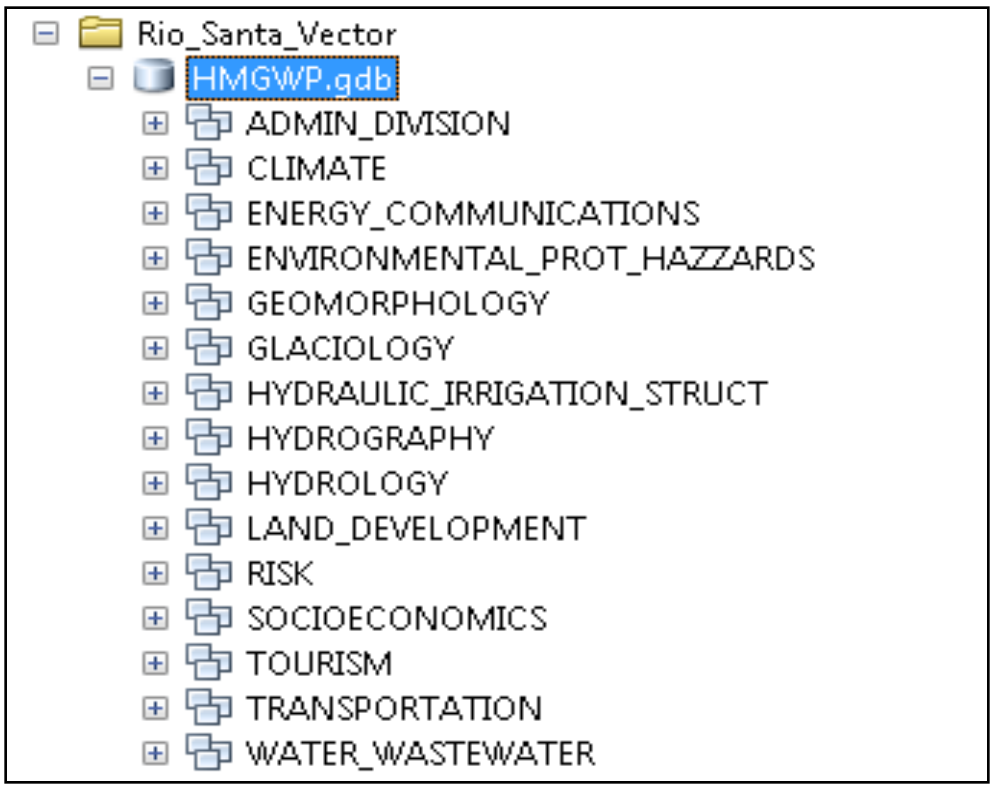

Figure 1. Image of the geodatabase organization structure.

Task 3: Populate the Geodatabase

Once I had received data from the McKinney group and the Peruvian team and settled on an organization structure I had to import all datasets to the Rio Santa geodatabase. Importing the datasets to the geodatabase using Arc Catalog converted all datasets into the same coordinate system. The geodatabase is in the WGS_1984 coordinate system and the Transverse Mercator projection. Although we have not yet discussed with the community of practice whether this coordinate system and projection is appropriate, we can easily make changes to these settings now that all datasets are in one geodatabase. Figure 2 shows the map of the resulting geodatabase.

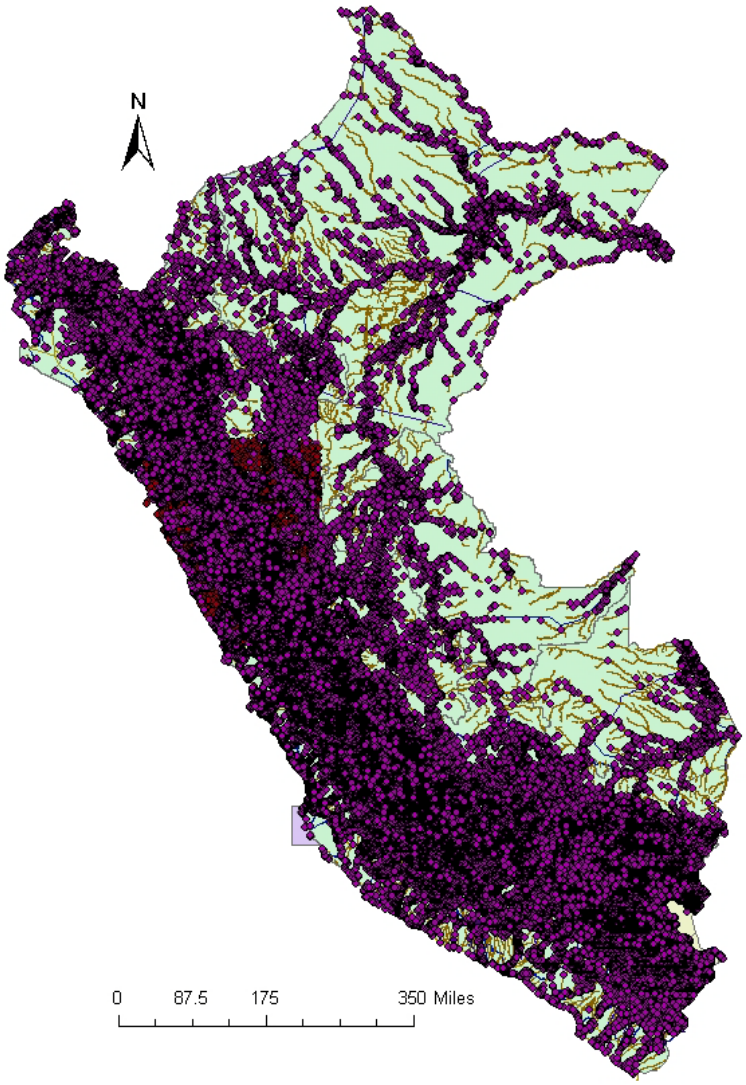

Figure 2. Map of the combined vector datasets from the McKinney group and Peruvian collaborators.

In combining data sets into the geodatabase organization structure I realized that rasters are not compatible with the structure we had created. Rasters cannot be added to the 'folders' (also called feature datasets) within a geodatabase. Therefore if I wanted to include them in the Rio Santa geodatabase they would not be grouped into categories in the organization structure. Another problem I encountered was the large size of raster datasets. By combining them with the vector files the geodatabase would be cumbersome to download and manipulate. Therefore I decided to create two geodatabases, one for raster datasets and the other for vector datasets. Although the raster geodatabase will not be organized in feature datasets like in the vector geodatabase, with proper naming of the rasters the geodatabase will be easy to navigate.

Task 4: Create New Layers

My original goal was to convert the tabular survey results into geographically referenced vectors displaying demographic and risk perception results. Nonetheless, the datasets we received from the Peruvian team were aggregated into rural and urban categories and had little geographic information for the survey sites. Given the lack of geographic information and low geographic resolution of the survey results, I did not create a social data layer for this project.

I also found that the geodatabase lacked information on ground cover throughout the Rio Santa basin. Information about ground cover is important for modeling how a flood will extend through a region. Therefore for this task I used Landsat images over the Western coast of Peru to calculate the NDVI in the Rio Santa basin. I also used two different categorizing tools in ArcGIS to try and identify what ranges in the NDVI results correspond to different kinds of ground cover.

Three image tiles from Landsat were needed to cover the basin area. The images I downloaded from the Landsat database are from before 2003. I did not use more recent data because of damage to the satellite imagining hardware that occurred in 2003 (NASA, 2012), which produced black strips on one of the tiles covering the Rio Santa basin. The images with the least cloud coverage were taken between 2000 and 2002. Figure 3 shows the images obtained from Landsat with the outline of the Rio Santa basin in yellow.

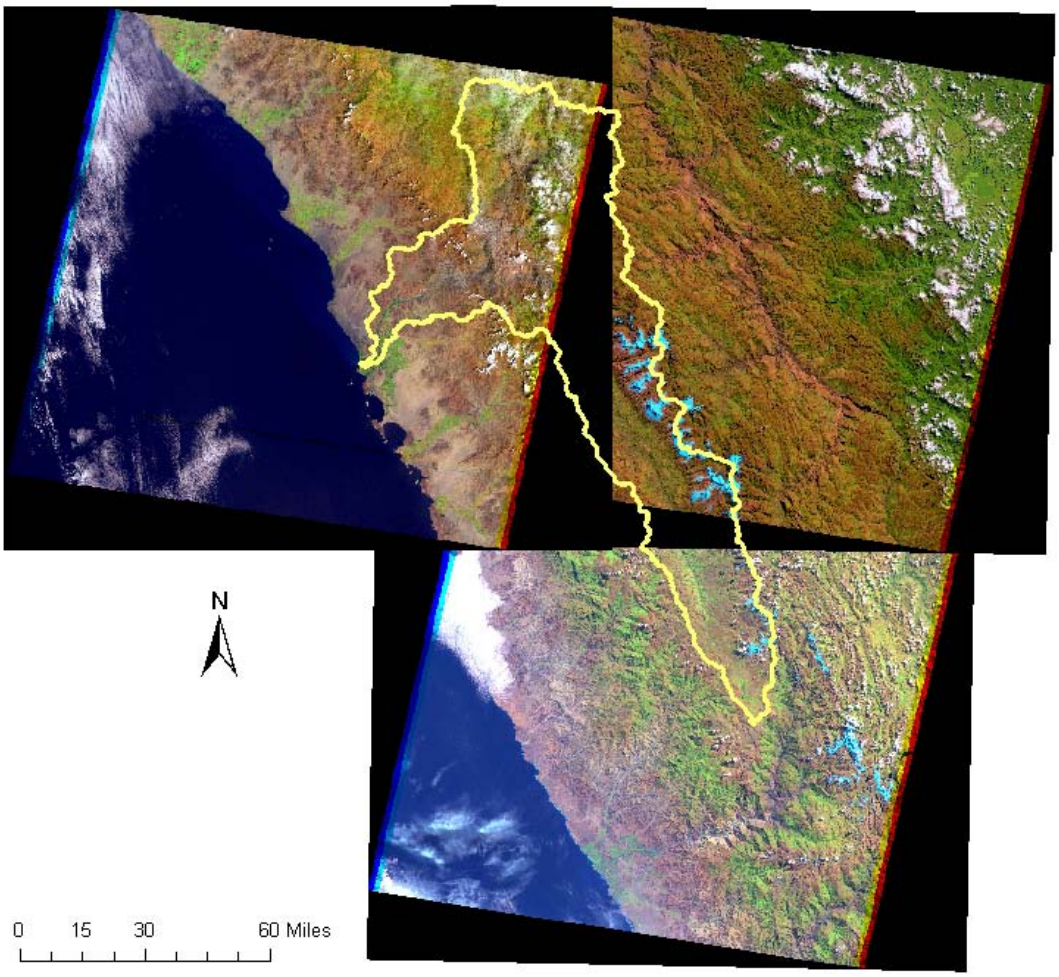

Figure 3. Landsat images used for the NDVI analysis with an overlay of the Rio Santa basin outline.

Next I calculated the NDVI of each tile individually using the NDVI tool in the Image Analysis window of ArcGIS. NDVI quantifies the relative difference between near infrared radiation reflected by the landscape and the red reflectance. Because green vegetation absorb red light, NDVI gives a measure of vegetation. The results of the NDVI calculation are shown below in Figure 4.

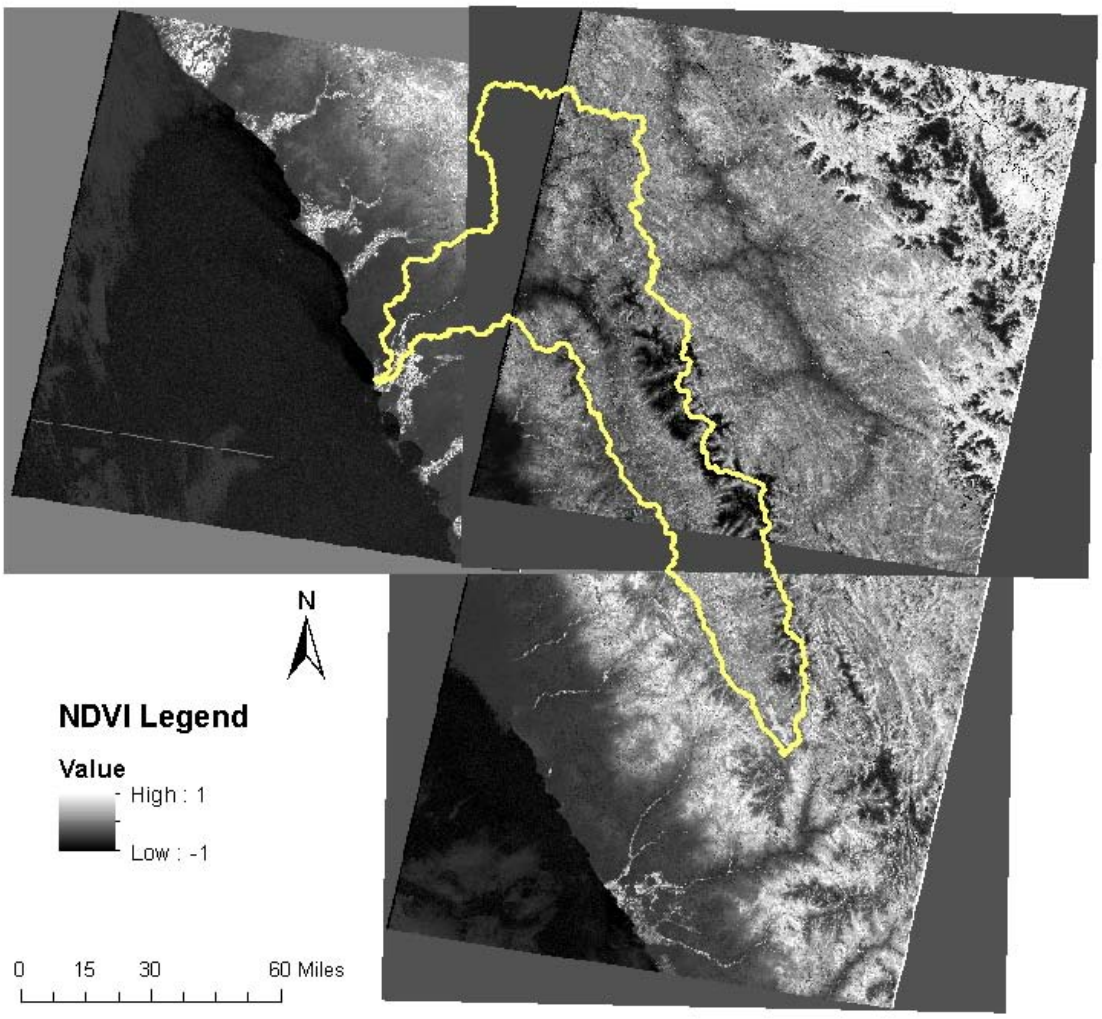

Figure 4. NDVI results for the Rio Santa Basin.

Before classifying the different NDVI values, I had to combine the three NDVI tiles using the Mosaic to New Raster tool. This tool decides which layer to retain in the mosaic raster by choosing the largest, smallest, last, first, mean or a blend of the overlapping rasters. Because the black borders around the mosaics had a value of zero and the minimum value for NDVI was -1, I had no way of indicating to ArcGIS that I wanted to retain the non-zero values. Therefore I used the Con tool to convert all zero values in each tile to -10. I expect that all NDVI values from the actual images are only close to zero and not exactly zero. Once the zero values were reclassified I was able to use the Mosaic to Raster tool so the tool would retain the raster with the greatest value wherever two tiles overlapped. This process produced the NDVI mosaic raster shown in Figure 5.

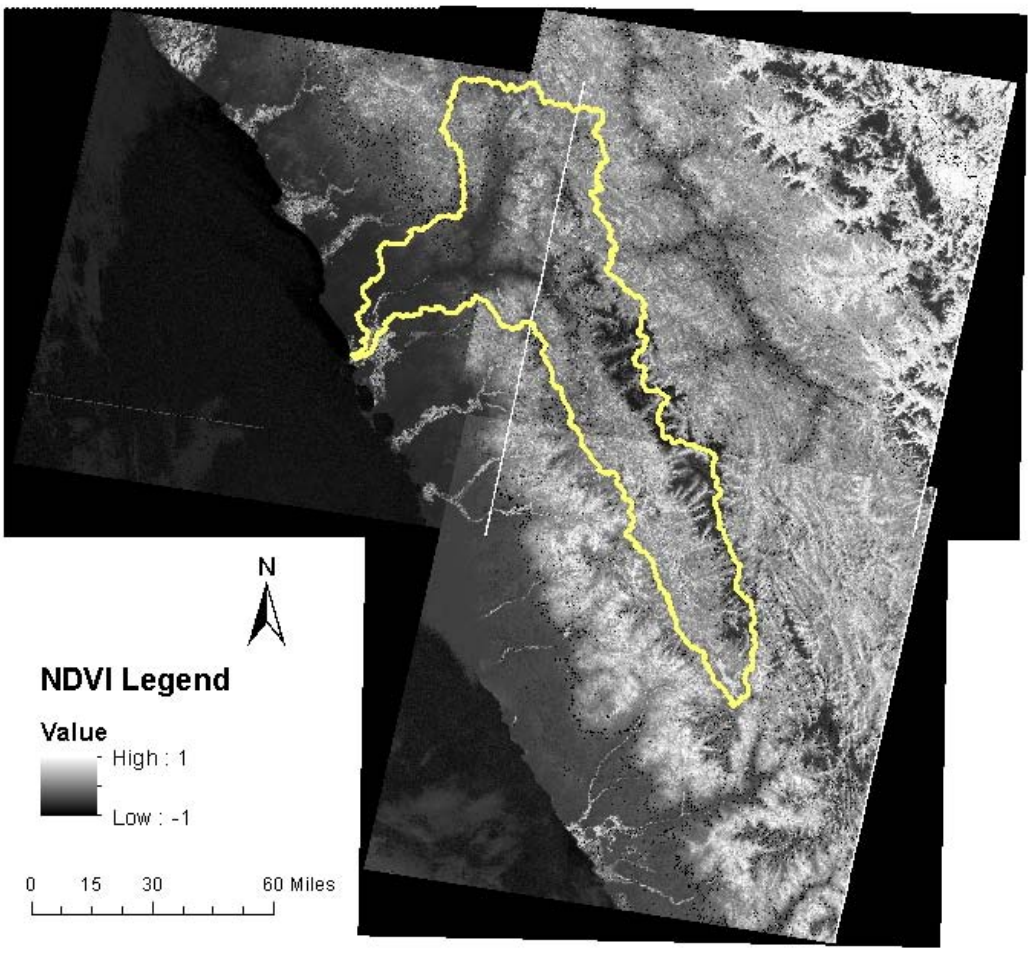

Figure 5. Landsat NDVI results combined into one raster.

The values of NDVI do not have a general physical meaning, but must be classified for the landscape and vegetation being studied. Data sets exist at various resolution levels and from various sources that provide land cover information for countries around the world (NASA, n.d.; USGS, 2010). Nonetheless many of the available datasets are not recent, therefore I chose to determine if the classification tools available in ArcGIS can classify NDVI values for the following categories: 1)dense vegetation 2)grass or pasture 3)bare soil 4)snow and ice 5)water. In addition to these categories I expect that there will be a category for the black bands around the edges of each tile and one for cloud cover. I chose not to trim the NDVI mosaic for the Rio Santa basin to include more land area and land cover types for the classification task. My hope is that by including examples of dense vegetation that will be more prevalent towards the eastern edge of Peru the classification tool will be able to distinguish between dense vegetation and grass lands with greater ease.

The first tool I used for the classification task was Reclassify. I used this tool to reclassify the NDVI mosaic raster using natural jumps in the data. This method assumes that there will be clear delineation between the different ground cover classes in the NDVI data. The results of creating 7 classifications using the Reclassify tool are shown below as Figure 6.

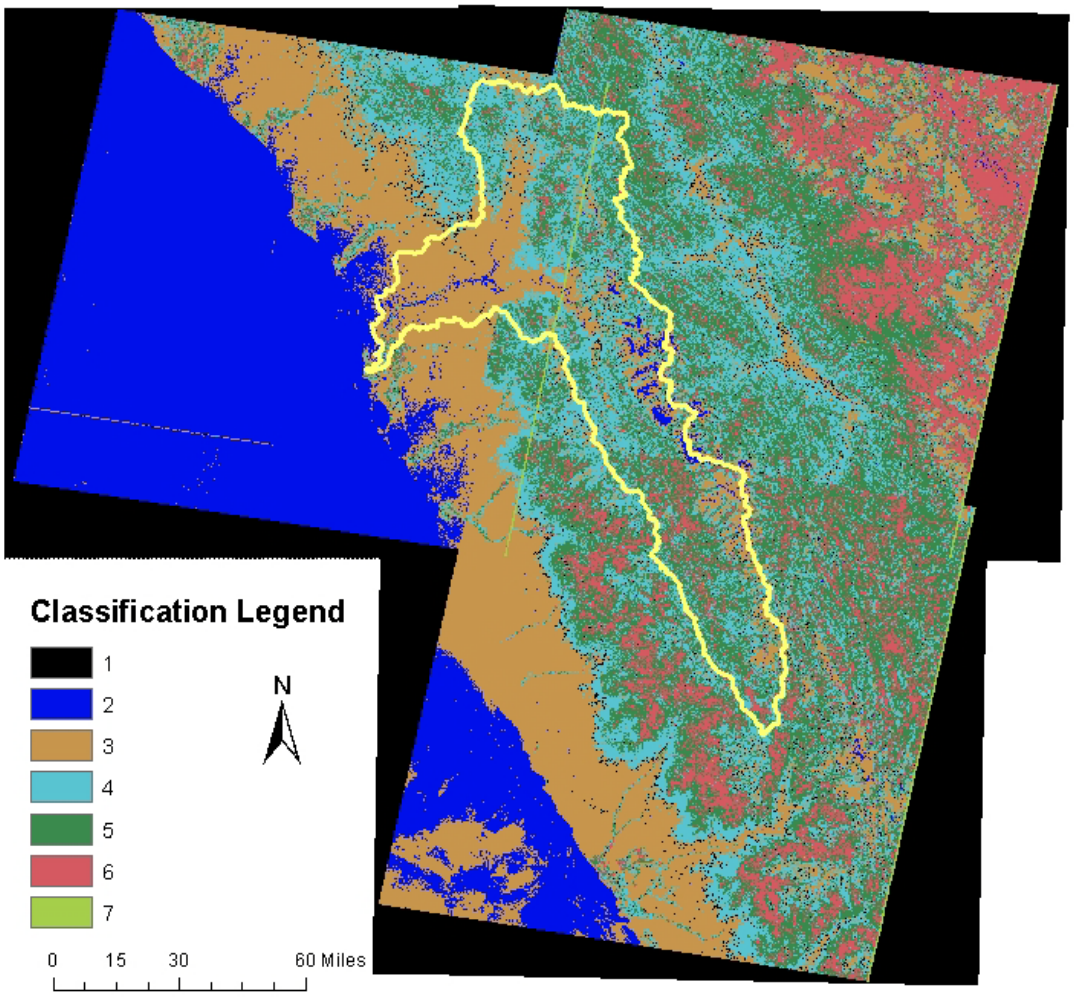

Figure 6. Results of classification using Reclassify tool.

Figure 6 shows that the Reclassify tool was successfully able to distinguish between water (classification 2) off the coast of Peru and land. There is some blurring of the coastline, though. Nonetheless the areas of apparently dense vegetation that can be seen in Figure 4 inland of the southern coast and in the North East corner of the mosaic are a muddle of classes 5, 6, and 7. Figure 7 shows a zoomed image of a section of the classified mosaic with the outline of glaciers (in white/light blue) and drainage lines. The vectors added to Figure 7 are from the vector geodatabase.

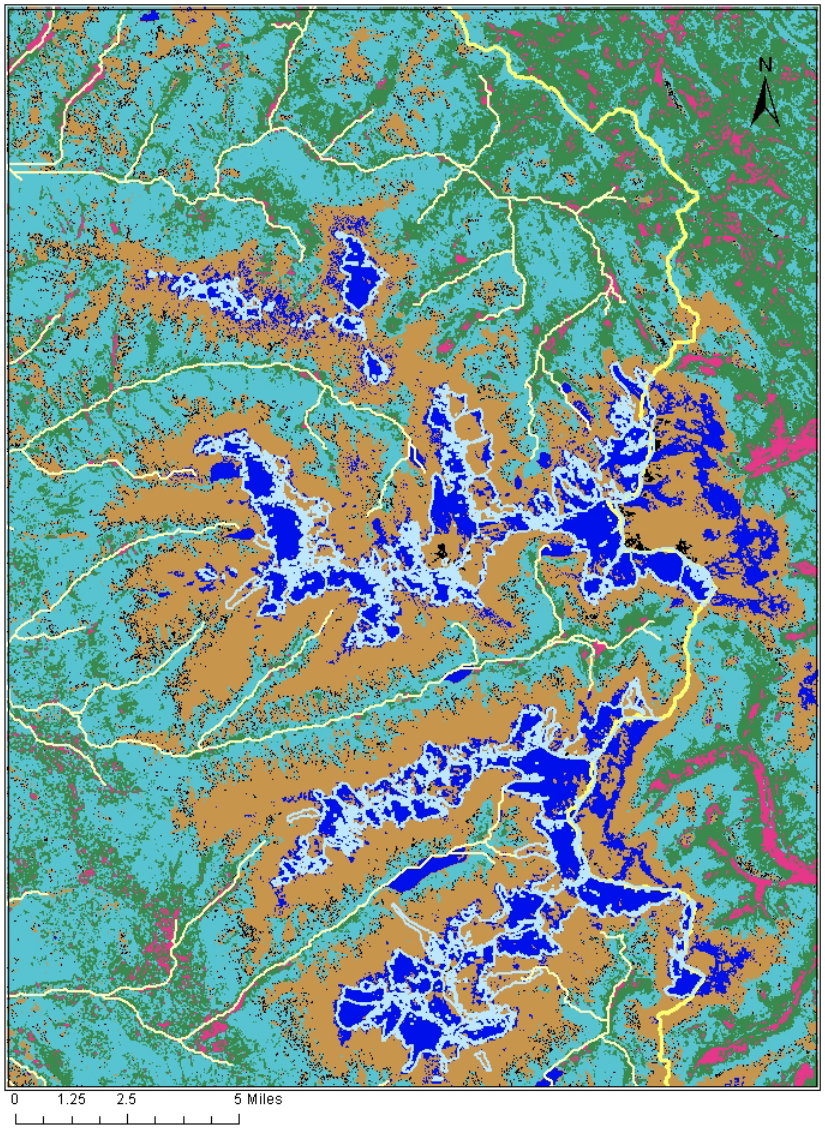

Figure 7. Close up of a region of the Rio Santa basin with drainage lines and glacier outlines.

Figure 7 shows that glaciers have been included in the water class and that drainage lines are not classified as water by this approach. Given that the Landsat images have a 15 m resolution, they may not have the detail to sense narrow rivers. Using the same zoom setting and layering the drainage lines and glacier outlines over the landsat image gives Figure 8.

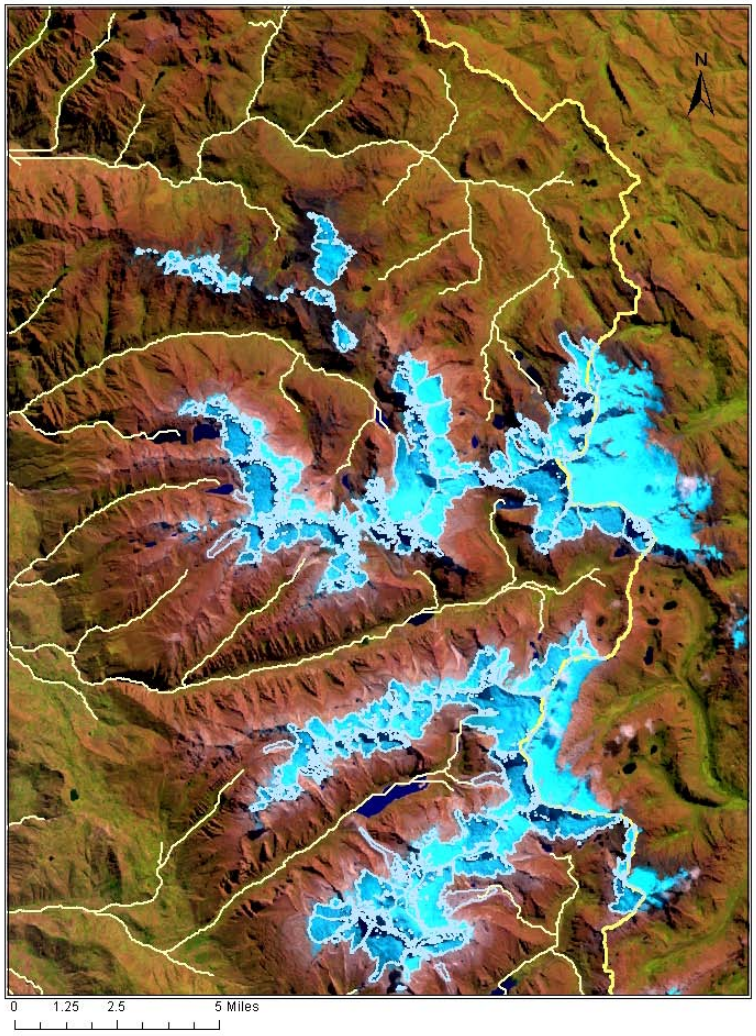

Figure 8. Landsat image with glacier outlines and drainage lines.

Comparing Figure 8 to Figure 7 suggests that class 3 represents bare soil and classes 5 and 6 represent vegetation. Class 4 appears to coincide with shaded areas of the image, which also appear to have some vegetation cover. Given that class 7 has the highest NDVI values, I assume this class also represents vegetation. The NDVI ranges for each class are summarized in Table 1.

Class number	NDVI range	Possible land cover
1	-1 to -0.64	Tile edges
2	-0.64 to -0.2	Water/ice
3	-0.2 to -0.067	Bare soil
4	-0.067 to 0.106	Shaded surfaces/vegetation
5	0.106 to 0.325	Vegetation
6	0.325 to 0.718	Vegetation
7	0.718 to 1	Vegetation

NASA Earth Observatory and Holben give some broad ranges of NDVI and what land cover they indicate (Holben, 1986; Weier & Herring, n.d.). These values are summarized in Table 2.

Table 2. Literature value for NDVI range and ground cover (Holben, 1986; Weier & Herring, n.d.).

NDVI Range Holben	NDVI Range NASA	Ground Cover
-1 to -0.257	<0.1	Water
-0.257 to -0.046	<0.1	Snow and ice
-0.046 to 0.002	No value	Clouds
0.002 to 0.025	<0.1	Bare soil
0. 025 to 0.14	0.2 to 0.3	Shrub and grassland
0.14 to 0.5	0.6 to 0.8	Dense vegetation

Because the NASA ranges are so large, they show better agreement with the classification results in Table 1 than the Holben ranges. The NASA and Holben ranges are general ranges for world land cover. Therefore they are not calibrated for the soil and vegetation types that exist in Peru and should not be expected to show close agreement with Peru specific data. I was unable to find specific NDVI values and land cover for the Western coast and Rio Santa basin of Peru.

The second classification tool I used was Isocluster Unsupervised Classification. This tool uses statistical methods to determine if a given NDVI value is likely to fit in the same class as another value. Figure 9 shows the results of the classification from the Isocluster tool.

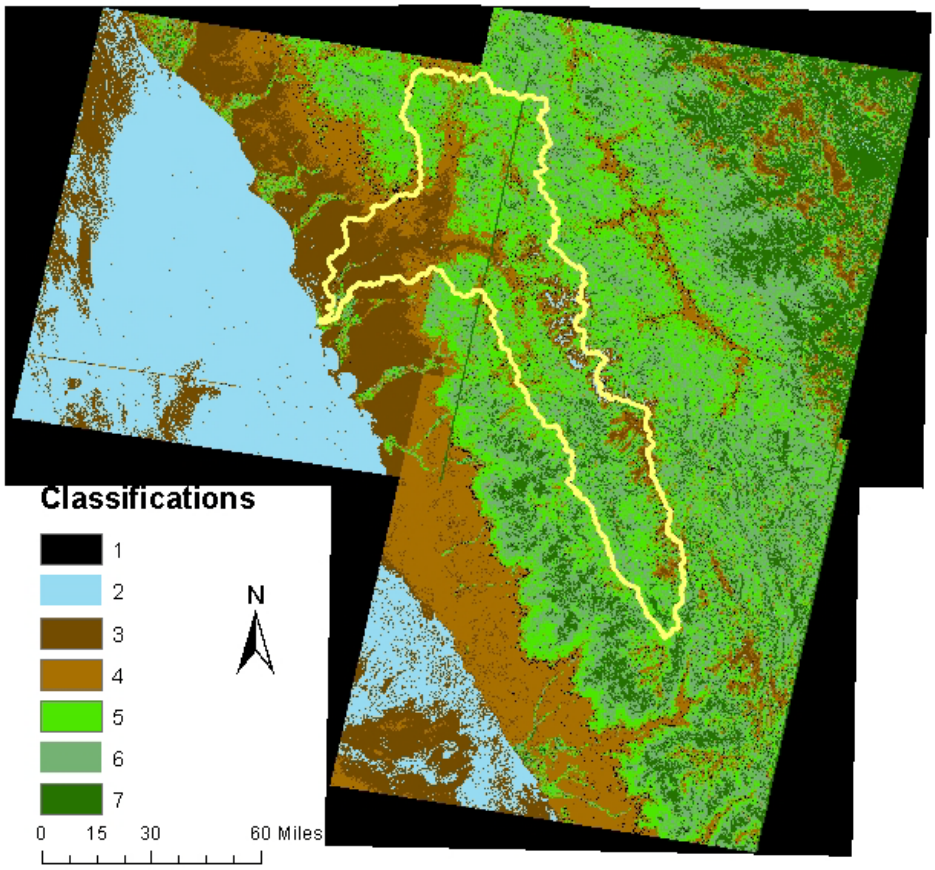

Figure 9. Results of the Isocluster Unsupervised classification tool.

In Figure 9 the coastline shows good differentiation between the ocean (class 2) and the shoreline that does not have the mixing effect seen in Figure 6. Nonetheless, clouds offshore are in classes 3 and 4, which also include bare soil. Offshore clouds were also classified with class 3 in Figure 6, but to a lesser extent.

In Figure 10 I show a higher zoom image including drainage lines and glacier outlines.

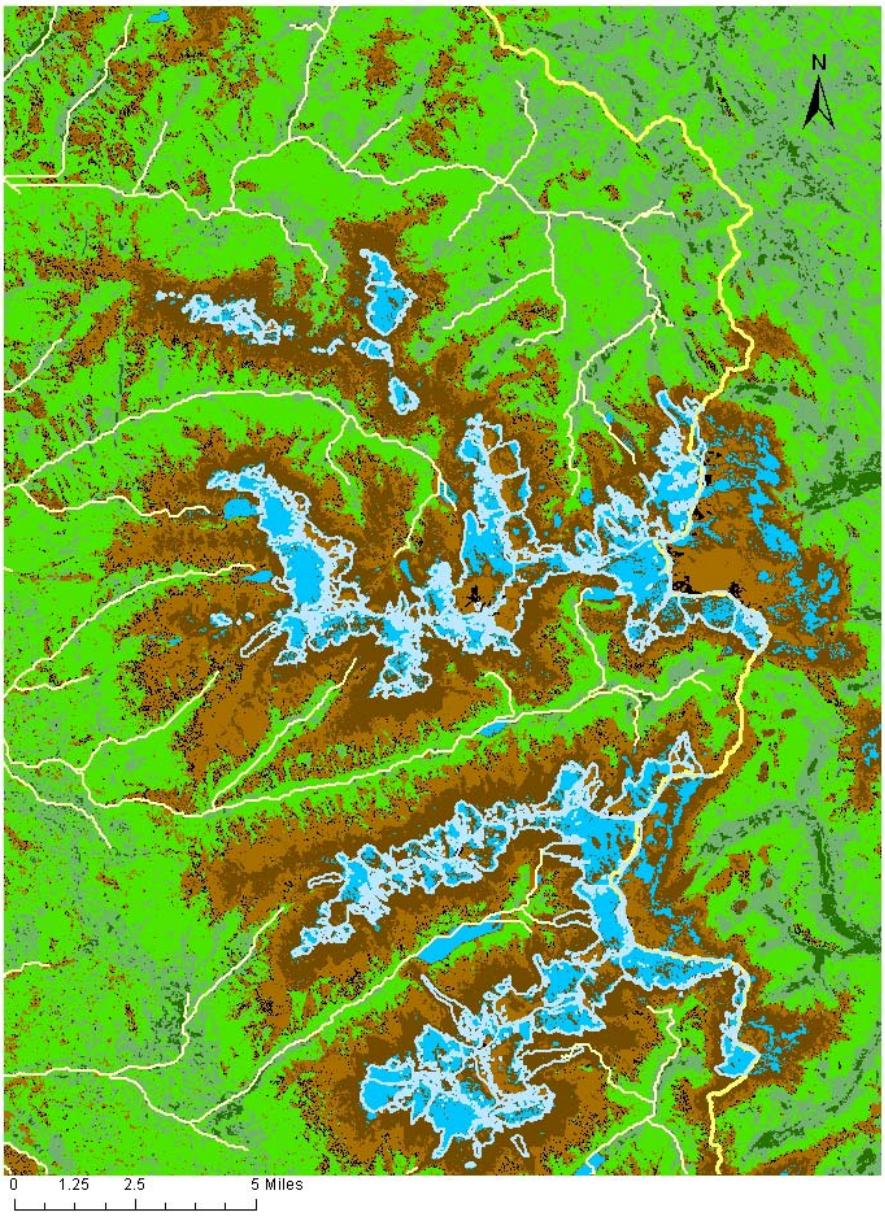

Figure 10. Higher zoom image of the Isocluster Unsupervised classification with glacier outlines and drainage lines.

Much like the previous classification, rivers are not included in class 2 in the Isocluster Unsupervised classification. Also, the glacier outlines show good agreement with class 2 areas in Figure 10 indicating that the classification method was able to distinguish water and ice from other ground cover.

Figure 10 can be compared with Figure 8 since they are images of the same region of the Landsat image and map. This comparison shows that the Isocluster Unsupervised classification correctly classifies bare soil and does not classify shaded vegetation regions separately from vegetation and bare soil as was seen from the Reclassify tool results. Instead the Unsupervised classification

results appear to have two bare soil classes (class 3 and 4). Class 4 surrounds the edges of glaciers that appear lighter in Figure 8 and may have light snow cover or be frozen. The Isocluster Unsupervised classification does not output the break points for each classification, therefore Table 2 only contains the classes and possible land cover category.

Class	Possible land cover
1	Tile edge
2	Water/ice
3	Frozen bare soil/light snow
4	Bare soil
5	Vegetation
6	Vegetation
7	Vegetation

Neither classification method distinguished water from ice. Nonetheless, the use of a vegetation index does not take into consideration the reflectance differences of ice and water. Overall the Isocluster Unsupervised classification tool was better able to distinguish the shoreline from ocean water and corresponded well with the glacier outlines obtained from researchers. In addition this classification tool distinguished two categories of bare soil.

Task 5: Share the Geodatabase

My initial goal was to share the Rio Santa basin geodatabase via Arc GIS online. Sharing a map package via Arc GIS online, though has several shortcomings. Uploading the map package is slow, the recipient must have Arc GIS to view the map, and downloading the map package will also be time consuming. Given that many of the researchers in Peru have intermittent and low bandwidth internet connections and few have Arc GIS, the map package sharing mechanism was inappropriate. Arc GIS online also has a platform for sharing maps where they can be viewed in an internet browser. Nonetheless the limitations on map size, functionality, and the ability of users to upload information also made it inappropriate for sharing the Rio Santa Geodatabase.

Task 5 remains incomplete, but is a topic I will work to address as the Geodatabase develops.

Conclusion and Future Work

Overall I found Arc GIS to be a good medium to gather, organize and present data from the Rio Santa basin. Considerations for the internet capabilities and proficiency with Arc GIS of our collaborators and consultation with the McKinney group has led me to amend my expectations for the geodatabase sharing platform. I now think that a better model for sharing the Geodatabase is to have one 'lightweight' (in terms of bandwidth requirements) platform that can be fully displayed in a web browser for users that wish to visualize and explore available data sets. For users that would

like to download and manipulate data sets in Arc GIS, we plan to create a visualization platform that allows for dataset downloads and uploads. Creating a platform with these capabilities will be addressed in future work in conjunction with the McKinney group and collaborators (hopefully!) from the Maidment group.

In creating the land cover classes I found that Arc GIS was an excellent platform to import and process Landsat images. I think these images can be very useful for those studying glaciers, because they show how glaciers evolve over time. Nonetheless the damage to the Landsat 7 imaging hardware limits the usefulness of the images in recent years.

In terms of creating land cover classifications, I found that the Isoclast Unsupervised classification tool results provided a better match to the land cover observed in Landsat images than the Reclassify tool using natural breaks in the data. Nonetheless I think more work is needed specifically for Peru to determine land cover classes for NDVI ranges.

Finally, I would like to work with the Peruvian researchers gathering social data to obtain better resolved data sets. The Peruvian team is very excited to be able to compare perceptions of risk with potential flood areas. I will be working closely with the Peruvian team in the coming months and hope to produce social data layers for the geodatabase.

References

Holben, B. N. (1986). International Journal of Remote Sensing Characteristics of maximum-value composite images from temporal AVHRR data. *International Journal of Remote Sensing*, *7*(11), 37–41.

NASA. (n.d.). Global Landuse Datasets. *National Aeronautics and Space Administration: Goddard Institute for Space Studies*. Retrieved from http://data.giss.nasa.gov/landuse/

NASA. (2012). Landsat 7. *National Aeronautics and Space Administration*. Retrieved from http://landsat.gsfc.nasa.gov/about/landsat7.html

USGS. (2010). South American Landcover Data Links. *USGS Land Cover Institute*. Retrieved from http://landcover.usgs.gov/landcoverdata.php#sa

Weier, J., & Herring, D. (n.d.). Measuring Vegetation (NDVI & EVI). *NASA Earth Observatory*. Retrieved from http://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_1.php