

University of Pittsburgh

Using Statistical Analysis to Determine Geographic Relationships
Among Automatic Mapping Errors in the Amazon Basin

Tom Plazek

GEOL 1901

1 April 2023

Abstract

This analysis and discussion section presents the results of hot spot analysis and spatial autocorrelation using the Getis-Ord G_i^* statistic and Global Morans I index, respectively, to identify patterns and relationships in river centerline data errors. The hot spot analysis confirmed the presence of hot spots with a high level of confidence for errors related to 'not a river' and 'oxbow lakes'. The missing centerline errors were also analyzed, revealing hot spots in the northwest corner and central region of the region of interest. The results were consistent with the overall error density map. Spatial autocorrelation analysis showed a significant spatial relationship between centerlines ending abruptly and river width, as well as a strong correlation between missing centerlines and river width. The findings highlight the importance of considering various factors, such as river width, in remote sensing analyses. However, limitations due to small dataset size and other factors, such as cloud and tree cover, should also be taken into account. Overall, this analysis provides insights into the spatial patterns and relationships in river centerline data errors, contributing to a better understanding of the data quality and potential sources of errors in remote sensing analyses.

Table of Contents

Abstract	2
List of Figures	4
1. Introduction.....	6
1.2 Background.....	7
2. Literature Review.....	8
2.1 Optimized Hot Spot Analysis with Hierarchical Density Based Spatial Clustering	8
2.2 Spatial Autocorrelation using the Moran's I Statistic	9
2.3 Hierarchical Density Based Spatial Clustering (HDBSCAN)	9
2.4 Empirical Bayesian Kriging (EBK).....	9
3. Methods.....	10
3.1 Background on Data and Data Acquisition	10
Figure 1 - The Amazon Basin with its rivers shown in white. 11	
3.2 Creating New Error Feature Classes.....	12
3.3 Optimized Hotspot Analysis (OHS)	12
3.4 Spatial Autocorrelation Analysis	13
3.5 Hierarchical-Density Based Spatial Clustering (HDBSCAN).....	13
3.6 Empirical Bayesian Kriging.....	14
Figure 2 - Initial map setup and feature class creation workflow.....	15
4. Analysis and Discussion	15
4.1 Hot Spot Analysis using Getis-Ord G_i^* Statistic.....	15
Eq. 1 - Getis-Ord G_i^* equation16	
Figure 3 - Optimized hotspot analysis for all errors combined	17
Figure 4 - Optimized hot spot analysis for 'not a river' error	18
Figure 5 - Optimized hot spot analysis for 'Oxbow lakes' error	19
Figure 6 - Optimized hot spot analysis for 'Missing centerline portions' error.....	20
Figure 7 - Optimized hot spot analysis for dam cluster	21
4.2 Spatial Autocorrelation using Global Morans I	22
Eq. 2 - Global Moran's I z-score equation	22
4.3 Hierarchical Density Based Spatial Clustering.....	24
Figure 9 - HDBSCAN Clustering for all errors combined25	

Figure 10 - Hierarchical clustering dendrogram of the HDBSCAN clusters for all errors..	26
4.4 Empirical Bayesian Kriging.....	26
Eq. 3 - General kriging equation.....	27
Figure 11 - Empirical Bayesian Kriging surface based on error ID	28
Figure 12 - Empirical Bayesian Kriging error surface based on error ID	29
5. Conclusion	30
<i>Special thanks</i>	30
Appendix A: Source Code	31
R- Source code for hierarchical cluster analysis and dendrogram.....	31
Appendix B: Supporting workflows	32
Figure 13 - OHS analysis workflow	32
Figure 14 - Spatial autocorrelation workflow.....	32
Figure 15 - HDBSCAN workflow	33
Figure 16 - EBK workflow	33
Works Cited	34

List of Figures

Figure 1 - The Amazon Basin.....	Error! Bookmark not defined.
Figure 2 - Initial map setup and feature class creation workflow. It should be noted that several of the steps were repeated for accuracy. This is what the duplicate arrows indicate.	15
Figure 3 - Optimized hotspot analysis for all errors combined	17
Figure 4 - Optimized hot spot analysis for 'not a river' error.....	18
Figure 5 - Optimized hot spot analysis for 'Oxbow lakes' error	19
Figure 6 - Optimized hot spot analysis for 'Missing centerline portions' error.....	20
Figure 7 - Optimized hot spot analysis for dam cluster	21
Figure 8 - Stream Order Count	Error! Bookmark not defined.
Figure 9 - HDBSCAN Clustering for all errors combined	25
Figure 10 - Hierarchical clustering dendrogram of the HDBSCAN clusters for all errors. A dendrogram is a ‘tree-like’ plot which shows, in this case, a correlation between the distance and the hierarchy of cluster values. Closer distances tend to correlate with higher clustering values.	26
Figure 11 - Empirical Bayesian Kriging surface based on error ID	28
Figure 12 - Empirical Bayesian Kriging error surface based on error ID	29

Figure 13 - OHS analysis workflow	32
Figure 14 - Spatial autocorrelation workflow	32
Figure 15 - HDBSCAN workflow	33
Figure 16 - EBK workflow	33

1. Introduction

The Amazon Basin is one of the most biodiverse places on the planet. While the exact number of species is unknown, the basin is thought to be home between 50,000 to 80,000 plant species and 2.5 to 5 million animal species. In addition to being a bastion of biodiversity, the Amazon rainforest is one of the largest carbon sinks in the world. It is impossible to determine the exact amount of CO₂ that the rainforests remove from the atmosphere yet, estimates suggest that nearly two billion tons of carbon dioxide (CO₂) are removed each year. The Amazonian rainforests are of crucial importance to regulate the carbon cycle on the planet and help to regulate other planetary systems. More research needs to be adopted due to increase focus on learning about deforestation and its effects. Increases in human deforestation in the Amazon have been the cause of increasing concern, especially as climate change and the focus on carbon dioxide emissions become more pressing. Being able to differentiate between deforestation, which is caused naturally, versus anthropogenic deforestation is an important and developing field of research.

Remote sensing is used to identify areas in the Amazon Basin that are being deforested. Rivers in the Amazon basin are very dynamic, meaning that their centerlines change from year to year due to meandering. This meandering can greatly impact bank erosion rates, leading to the rivers' positions shifting greatly over relatively short periods of time. This meandering motion is common in the Amazon because of the large amount of cohesion caused by roots from land plants. The high drainage densities in the basin also impact this rate of centerline change. High drainage densities are common in the basin because it is home to the largest river in the world, the Amazon River. Many of its branches and tributaries are nearly as large.

Using automatic mapping methods on remotely sensed imagery (Landsat 1 -8) in the basin has allowed for areas which are water to be identified, as well as for their centerlines to be mapped, spanning back to the mid 1980s. Automatic mapping algorithms are not perfect though – errors can still be made. Areas such as oxbow lakes used to be part of river channels. Earlier remote sensing databases have these sections as being part of an active fluvial system, meaning that at one point in time they would have had centerlines. This can lead to errors in accurately identifying actively flowing bodies.

The aim of this research was to manually find and determine the errors created by the automatic mapping process, as well as determine if there was any spatial or relational significance to environmental variables.

1.2 Background

Once the errors were found and identified, statistical analyses were run on the datasets to determine if any areas in the region were most affected by automatic mapping errors. These analyses included using optimized hop spot analysis and high-density clustering to determine if there were any error hotspots and/or clusters, as well as their locations. Additionally, spatial autocorrelation using the Global Moran's I statistic was run on errors two and four. Lastly, a kriging surface, as well as a kriging error surface were created to predict areas of high error density. This was accomplished using Empirical Bayesian Kriging.

The five error classes which were identified in this research were as follows:

1. Areas which are not part of a river but were automatically identified as one.
2. Areas where the centerlines abruptly end.
3. Oxbow lakes which used to be part of a river but were cut off into separate bodies by meandering processes over time.
4. Areas where the river centerlines are missing in either small or large chunks.
5. Man-made reservoirs which have been automatically identified as river bodies.

Each of the five errors classes in this research will have areas where they are most prevalent. The errors are different in nature and are likely to be caused by different elements. Below are the proposed regions where each of the errors will be statistically significant:

1. *Not a river*: Many of the lakes in the middle of the basin will be mapped as rivers, due to their slow flows.
2. *Centerlines end*: The northwest region of the basin is where most of the headwaters originate for the rivers in the basin. This region is very mountainous, and the rivers are much narrower here. This region yields more abrupt ending of centerlines. Additionally, there are many regions in the middle of the basin which have downstream tributaries that narrow.

3. *Oxbow lakes*: Both the southern and middle regions of the basin have large sections of rivers which are heavily braided and meandering.
4. *Missing centerline areas*: Similar to the centerlines end error, the width of the river is thought to have a significant correlation with continuous centerlines. These errors in the mountainous northwest region, as well as small branch tributaries are anticipated to have statistically significant relationship to the width of the river.
5. *Reservoirs*: The reservoirs which have mistakenly been mapped as flowing bodies will most likely see their hot spots in the southern region, where deforestation and urbanization are the most prevalent in the region. These reservoirs are hypothesized to have spatially significant relationships to the location of dams in the region as well.

It should be noted that dams tend to originate around human settlements and activities. These dams can be used for several purposes, including hydroelectric power and reservoir creation.

2. Literature Review

2.1 Optimized Hot Spot Analysis with Hierarchical Density Based Spatial Clustering

Hotspots were compared to hierarchical clusters to confirm their statistical significance (Zerbe et al., 2022). This study used hierarchical clustering and optimized hot spot analysis to characterize spatial and temporal variation of large wildfires in Washington State. These fire datasets started in 1970 and ended in 2020. Large amounts of points were able to be analyzed using these two methods relatively easily. While their study included the temporal aspect, the statistical analysis of this study does not directly do so. Certain parameter choices were justified based on this study, including the choice to use a hexagon grid as opposed to a fishnet grid. The choice to arbitrarily set a cluster density for the hierarchical clusters calculated by Hierarchical Density Based Spatial Clustering (HDBSCAN) was influenced and justified by Zerbe et al. (2022). Using a hexagon grid reduces the distortion from the earth's curvature more efficiently than a fishnet grid. With the Amazon basin being over 6,000,000 km², it was especially important to preserve projection property and avoid distortion (Agarwadkar et al. 2013). Literature also reinforced the choice to use optimized hotspot analysis (OHS) analysis to determine whether hierarchical clusters followed the same trend as hotspots (Zerbe et al., 2022).

2.2 Spatial Autocorrelation using the Moran's I Statistic

Spatial autocorrelation is used in GIS to determine if proximity to a given feature with respect to another feature is statistically significant. These outputs give a p-value and z-score, both of which can be used to show statistical significance. They can be used in tandem or separately. Roberts et al. (2000) used spatial autocorrelation to determine if there was a geostatistical significance between polygons displaying woodland fragmentation. These fragments were sorted into large, medium and small sizes and analyzed in a number of ways, including clustering. Spatial autocorrelation was used to determine spatial relationships between the different polygons in the respective size frames which they were assigned to. This remains a useful tool in GIS for determining the significance of spatial relationships.

2.3 Hierarchical Density Based Spatial Clustering (HDBSCAN)

Hierarchical Density based Spatial Clustering can be broadly applied in GIS. It does have some downsides, as outlined Grubestic et al. (2001). These include the fact that HDBSCAN uses an arbitrary user input for what defines a meaningful cluster. What constitutes a 'meaningful' cluster must be chosen by the researcher within the context of the question being asked. This might be considered a shortcoming; however, this input allows researchers to have more control over their clustering analysis (How Density-Based Clustering Works—ArcGIS pro | Documentation, n.d.). Other authors, such as Zerbe et al. (2022), have used this technique to produce meaningful results in their study on wildfires in Washington State.

2.4 Empirical Bayesian Kriging (EBK)

EBK differs from traditional kriging by accounting for errors introduced from estimations in the semivariogram model (Krivoruchko & Gribov, 2019). A case study from Giustini et al. (2019) used EBK to create a prediction surface which used geochemical data taken from the area. This study used EBK to create a 'Geogenic Radon Potential' map which was used to assess and predict the severity of radon exposure to people living in the volcanic region in central Italy under study (Giustini et al., 2019). While this study predicted levels of radon to people living in the area, the use of EBK and the EBK regression model are broad. EBK can be used for any kind of spatial research which involves predicting the location of something based on pre-existing data. Using pre-existing data through Bayesian processes is what differentiates EBK from traditional kriging.

3. Methods

3.1 Background on Data and Data Acquisition

Pre-processed imagery of the Amazon basin was used for this research. The imagery was obtained from Dr. Elad Dente and Team at the University of Pittsburgh. This data was in the form of shapefiles and imagery, exported from Google Earth Engine. The imagery contained no underlying or supporting data but only displayed classified regions of deforestation caused by human activity. Width data, in the form of nodes, was downloaded independently from the Surface Water and Ocean Topography (SWOT) dataset, specifically the river data contained within the SWOT dataset (SWORD). Additionally, the imagery of the yearly centerlines (processed from LANDSAT datasets) was obtained. When running his analysis, Dr. Dente filtered out any rivers which had a width of under 120 meters. This was done to remove any rivers which could obscure data and create even more errors than necessary. It is difficult for satellites to accurately identify rivers smaller than 120 meters in width. The same was done for the SWORD dataset before joining it to the relevant attribute tables for errors two and four. Lastly, a dataset of known dams in the Amazon basin was given in the form of shapefiles. This imagery was inclusive of the entire basin, which spans roughly 6,000,000 km². Additional datasets were downloaded and added to the ArcGIS project independently. These included a dataset from Oakridge National Laboratory which provided detailed shapefiles for the basin area, a flow length grid, a flow accumulation grid and a flow direction grid. A digital elevation model (DEM) of the Amazon was used to confirm the general height of regions of interest in the basin. Height was not studied directly in relation to error densities.

The Amazon Basin

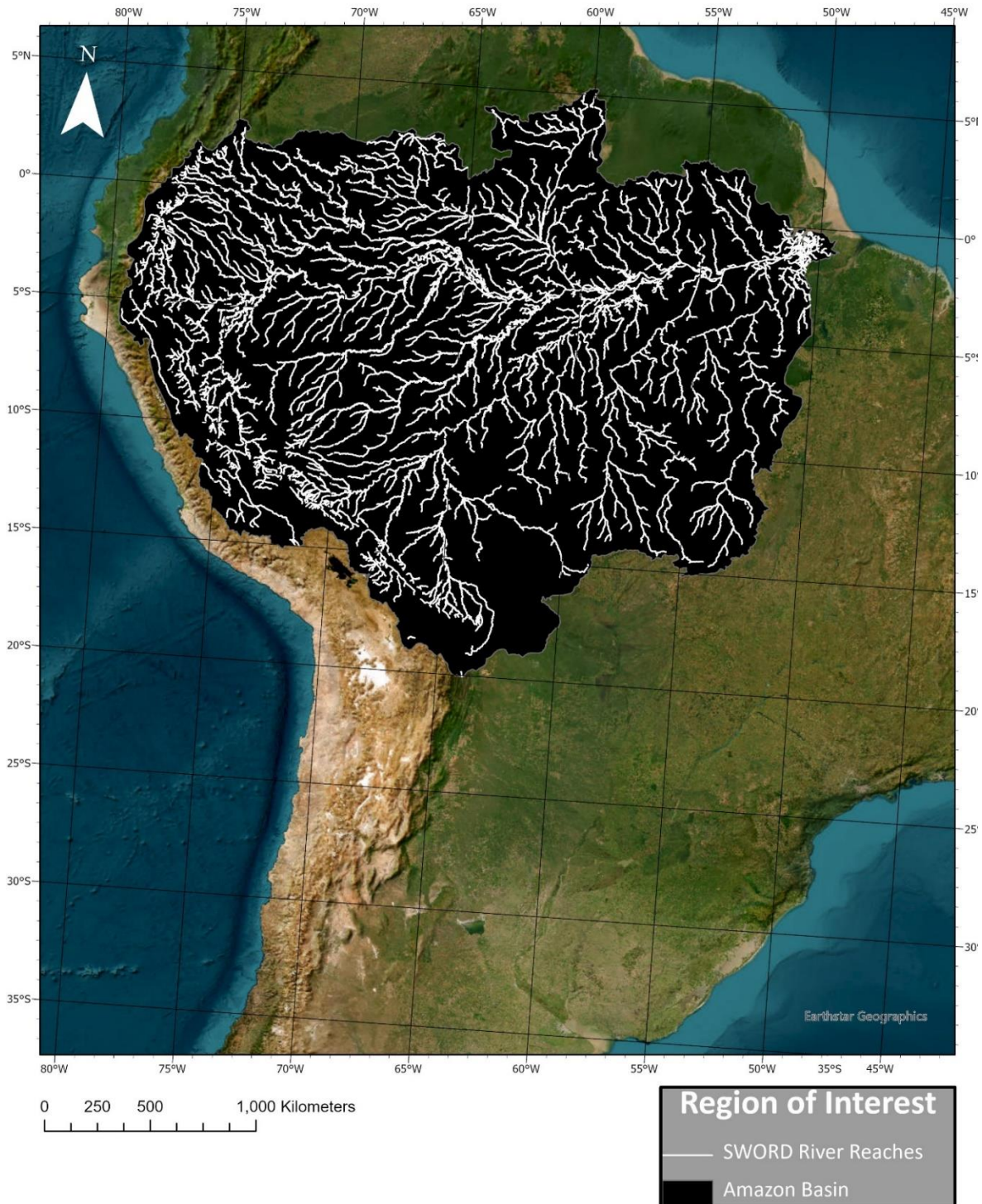


Figure 1 - The Amazon Basin with its rivers shown in white. The Amazon Basin is home to the largest river in the world, as well as an estimated 50,000 to 80,000 species. Its 6,000,000 km² create distinct challenges in studying this area. Dense leaf cover and complex topography, as well as yearly cloud cover are all sources of error when using remote sensing techniques to conduct research

3.2 Creating New Error Feature Classes

The first aim of this research was creating point and polygon feature classes to house the new error features. Second, the basin was systematically explored for errors visually. Originally only two feature classes were created using the ‘create feature class’ tool: one for error points, and one for the polygons. Once these feature classes were populated with points and polygons locating errors. using the ‘edit and add feature’ tools, a new column was populated with numeric indicators for each of the different kinds of errors. A number of one through five was assigned to the errors as they are listed above:

1. Areas which are not part of a river but were automatically identified as one.
2. Areas where the centerlines abruptly end.
3. Oxbow lakes which used to be part of a river but were cut off into separate bodies by meandering processes over time.
4. Areas where the river centerlines are missing in either small or large chunks.
5. Man-made reservoirs which have been automatically identified as river bodies.

Using the ‘select by attributes’ tool, separate feature classes were able to be created for the point feature classes of each error (using the ‘create feature from selection’ tool). Once these new feature classes were created, the points were all manually confirmed to be within their congruent polygons. This confirmation was simple done visually, so the ‘contains’ spatial join option could be used. The join was then run on each of the point feature classes to join them to the polygons. This simplified the data into one all-encompassing feature class for each of the errors.

3.3 Optimized Hotspot Analysis (OHS)

Once the error datasets were completed, statistical analysis was run to determine if there were any spatial relationships among the errors or their surrounding areas. First, optimized hot spot analysis was run to determine if there were any hot spots among the errors, as well as the dams (Fig.2). Optimized hot spot analysis used the Getis-Ord G_i^* statistic, which is a spatial autocorrelation measure that is used to identify and quantify the clustering of high or low values of a variable across a geographical region. The G_i^* statistic measures the spatial association between a particular location or point and its neighboring units with respect to a specific attribute or variable of interest. It then calculates the difference between the observed value and the expected value of the attribute for all neighboring units, taking into account the overall mean and variance of the attribute for the entire region. No specific analysis field was used for this research. A hexagon grid

mesh was chosen over a fishnet grid to better preserve spatial features and create more uniform outputs. Hexagon grids also reduce distortion from the earth's curvature more efficiently than a fishnet grid (Zerbe et al., 2022). Automatically calculated hexagon size was also kept for this research, since the OSH algorithm attempts to calculate the most optimal size for the inputted data. Changing these sizes would have been arbitrary. A bounding polygon was also used in this OHS analysis. This bounding polygon is the same one which was used in figure one and can be seen above. It creates a perfect outline of the Amazon basin. The remaining three error classes, as well as the dam feature class had optimized hot spot analysis run on them (Fig. 3, 4, 5, 6).

3.4 Spatial Autocorrelation Analysis

Once the statistically significant hot spots were established, spatial autocorrelation was run in relation to the river width. Only the errors which resided directly on the rivers could be spatially related to the width, therefore the not a river, oxbow lakes and reservoir datasets were excluded. This resulted in p and z-values which will be discussed below. It should be noted that in the original centerline dataset rivers which were under 120 meters in width were not included due to the difficulties in identifying them. This has caused some errors in mapping missing centerline portions, since some small portions of rivers under 120 meters may occasionally become 120 meters or wider, causing centerline segments to appear randomly.

3.5 Hierarchical-Density Based Spatial Clustering (HDBSCAN)

Hierarchical density clustering was run to determine the spatial location and significance of error clusters (Fig.7). This was also repeated for the dam dataset to compare the spatial significance of dam locations and dam-related errors. HDBSCAN is a clustering algorithm that can be used to identify spatial clusters based on the density of a feature dataset. A minimum cluster value can be set in the geoprocessing tool, so eight was chosen to be the minimum number of events to be considered a cluster. This was arbitrary due to the small size of the feature datasets. No specific literature was found on calculating minimum density feature numbers that were relevant to this research. Literature cited in the review above used HDBSCAN techniques and set arbitrary cluster values. Points considered to be outliers were excluded and the distance method used was Euclidian. The HDBSCAN clusters were visually compared to the relevant OHS output maps to see if the spatial patterns matched closely. Output maps had their symbology modified and changed

to a 'Jenks natural break', as opposed to a 'stretch'. Changing the symbology to a Jenks break allowed for distinct clustering regions to be formed, as opposed to a large spectrum.

Reservoirs, which are the error most closely linked to dams, were compared visually on a map. The cluster data was also exported from ArcGIS as a CSV and loaded into R. From there the clusters were parsed, removing any potential 'null' values from the dataset, bound into a scaled matrix (and plotted as a dendrogram (see Fig.9).

3.6 Empirical Bayesian Kriging

Empirical Bayesian Kriging (EBK) was used to create both a statistical prediction surface, along with an error surface for all errors combined (Fig. 8 and Fig. 9). The 'Geostatistical Wizard' in ArcGIS Pro was used to complete this task. The 'source data' was the combined error dataset, and the 'data field' used was the 'issue_id' field, which was the numerical identifier for each error type. EBK outputs were generated which indicated the spatial predictions for each type of error (i.e., errors one through five). The error surface was included due to the fact that ArcGIS Pro could not distinguish the fact that error ID codes were given in whole number intervals. Ranges in values were given on the kriging surface, however these should correlate with whole number identifies. The error surface was used to derive a better understanding of which whole number values aligned more closely to the prediction surface.

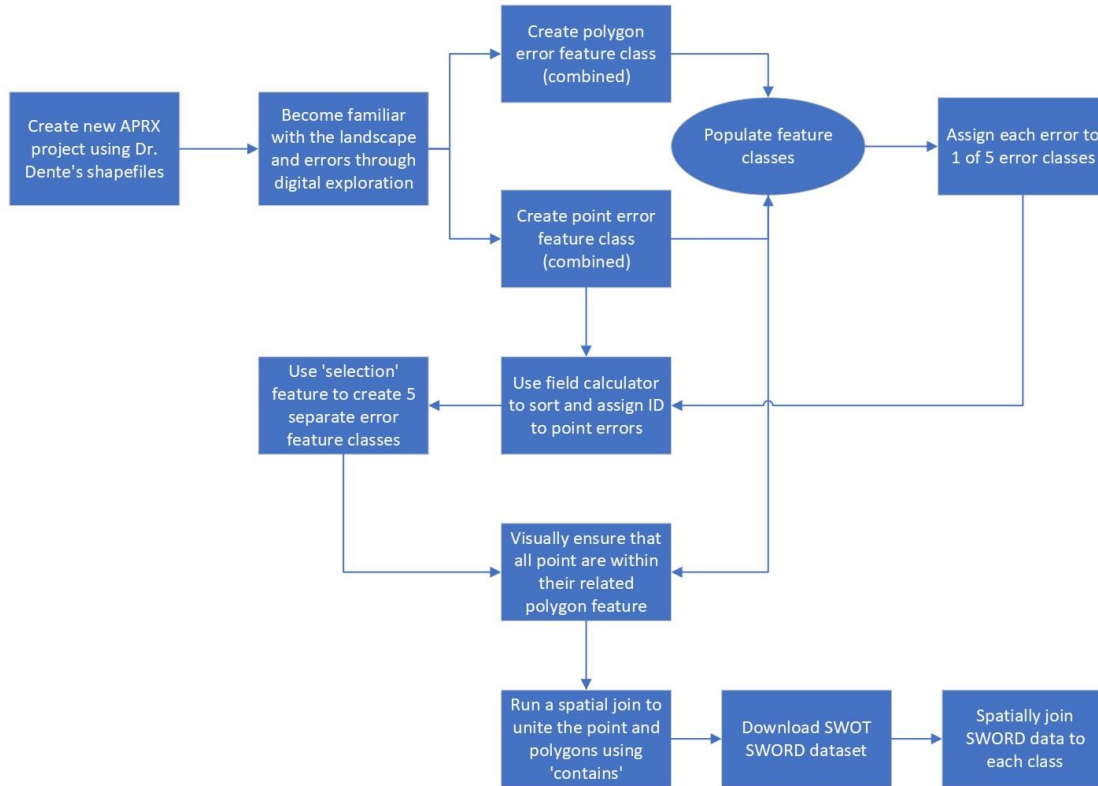


Figure 2 - Initial map setup and feature class creation workflow. It should be noted that several of the steps were repeated for accuracy. This is what the duplicate arrows indicate.

4. Analysis and Discussion

4.1 Hot Spot Analysis using Getis-Ord G_i^* Statistic

The hot spots which showed up for the ‘not a river’ error did occur where they were initially predicted to (see Fig. 3). Many lakes formed in the flatter, lower elevation offer floodplains that can make it easy for lakes to form over time with repeated flooding. While this area appeared to be a cluster visually, the Getis-Ord G_i^* statistic confirmed that it was a hotspot with 99 percent confidence. Similarly, the visual hot spot which appeared in the southern region for the ‘oxbow lakes’ error was confirmed by the optimized hot spot analysis (see Fig.4). This cluster was confirmed to be a hot spot with 99 percent confidence. The missing centerline errors had several hot spots – one in the northwest corner of the ROI and three in the central region toward the east. The results were more comprehensive, including confidence regions from ninety-nine down to ninety percent. Adjusting the unit size for the hexagons may be appropriate, due to the sheer size of the region. All these maps corresponded very closely with the overall error density map. The

‘centerlines end’ and ‘reservoir’ errors were not included separately, due to having too few points to individually calculate the optimized hot spots. A minimum of thirty independent points were required. The Getis-Ord G_i^* statistic which ERSI used in their optimized hotspot tool is calculated using the following equation:

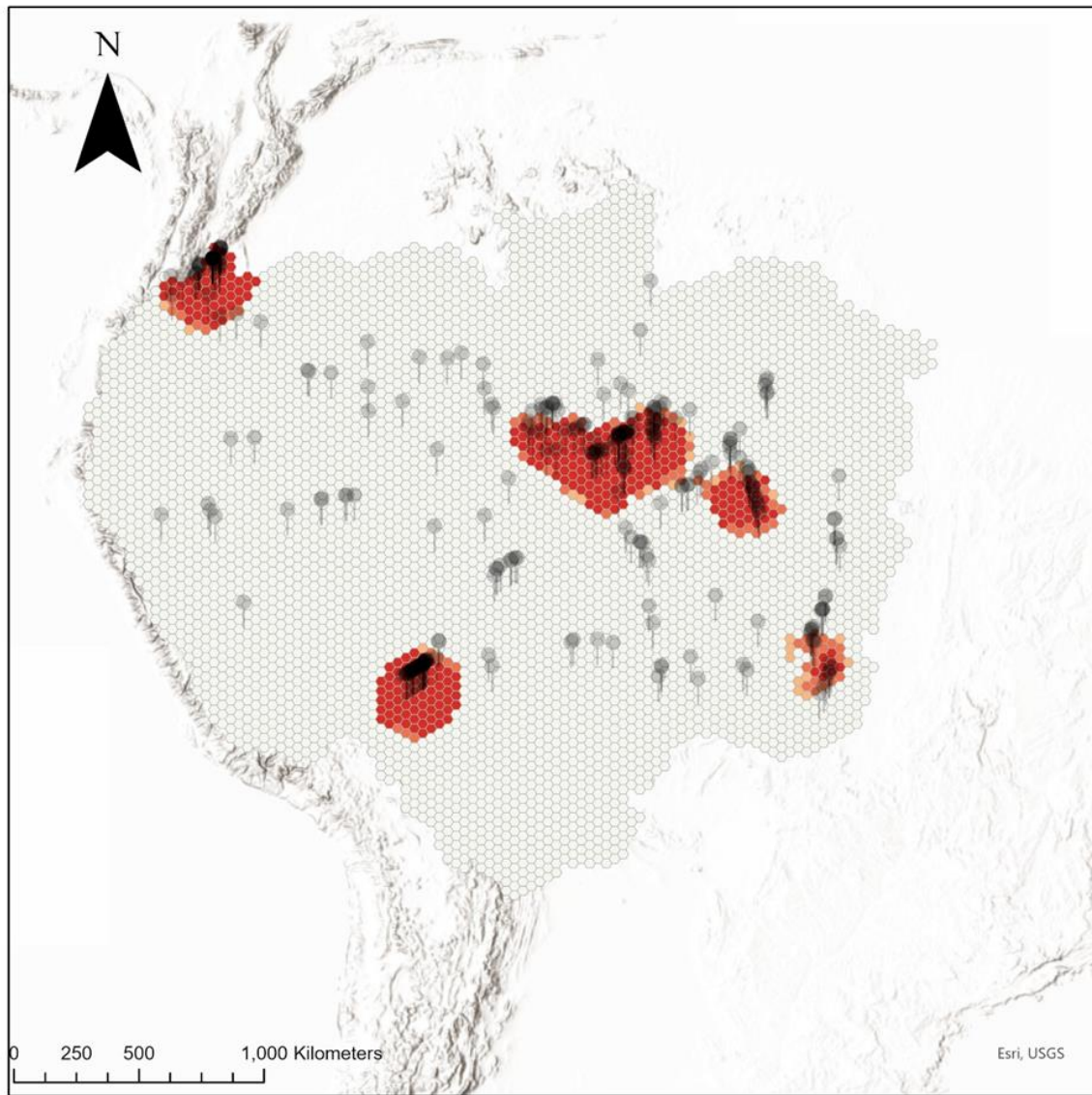
$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{(n \sum_{j=1}^n w_{i,j}^2 - (\sum_{j=1}^n w_{i,j})^2)}{n-1}}}$$

$$\bar{X} = \frac{(\sum_{j=1}^n x_j)}{n}$$

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{X})^2}$$

Eq. 1 - Getis-Ord G_i^* equation. In this equation, x_j is the attribute value for feature ‘j’, $w_{i,j}$ is the spatial weight between i and j and n is equal to the total number of features. \bar{X} and S are defined above. G_i^* returns a z-score, however the ‘Hot Spot Analysis’ tool calculates p-values as well.

Optimized Hotspot Analysis for all Errors Combined



Optimized Hotspot Analysis

Hot Spot Statistics

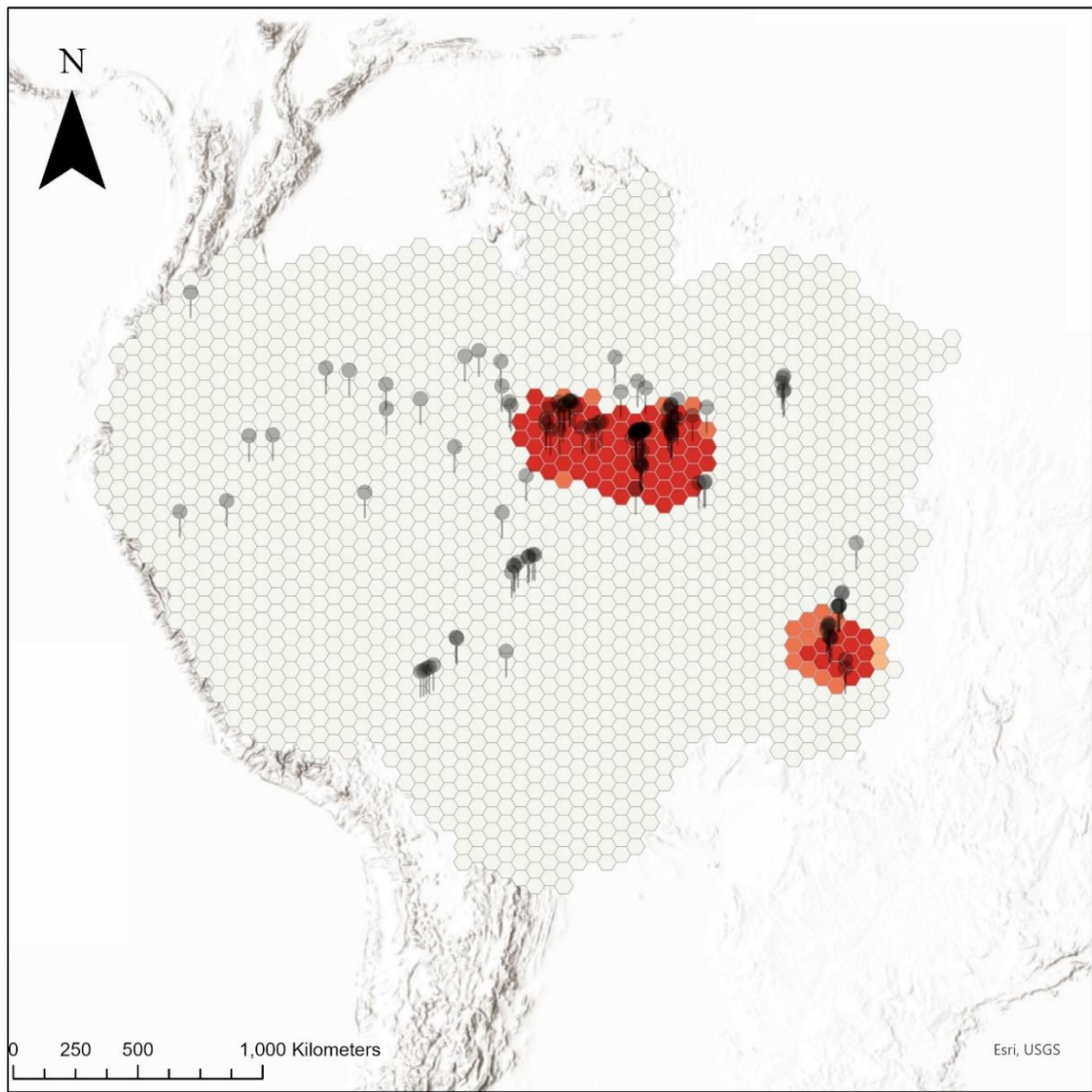
*Getis-Ord Gi**

- Not Significant
- Hot Spot with 90% Confidence
- Hot Spot with 95% Confidence
- Hot Spot with 99% Confidence

Combined Errors

Figure 3 - Optimized hotspot analysis for all errors combined

Optimized Hotspot Analysis for 'Not a river' Errors



Optimized Hotspot Analysis

Hot Spot Statistics

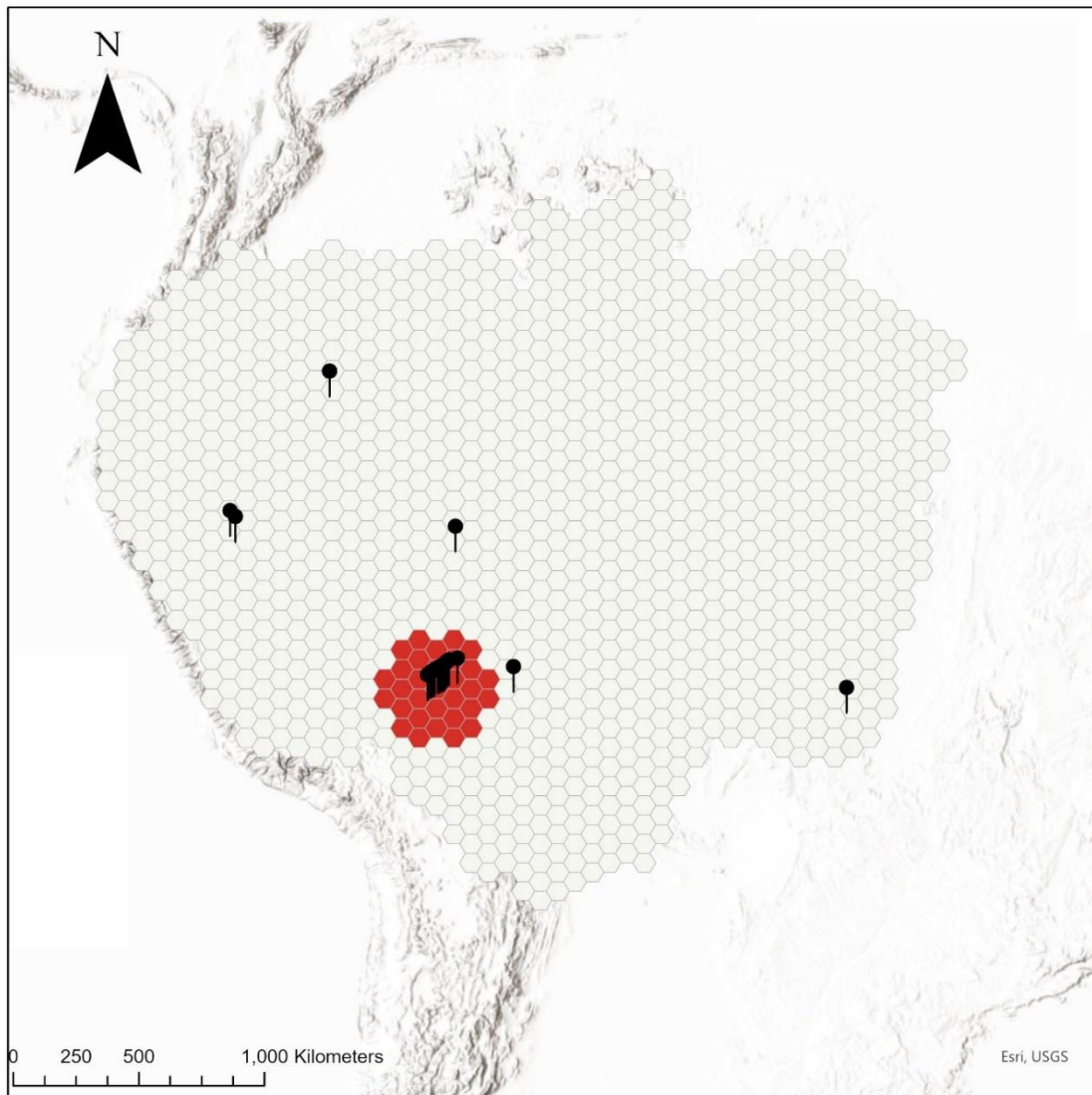
*Getis-Ord Gi**

- Not Significant
- Hot Spot with 90% Confidence
- Hot Spot with 95% Confidence
- Hot Spot with 99% Confidence

Not a river Error

Figure 4 - Optimized hot spot analysis for 'not a river' error





Optimized Hotspot Analysis for Oxbow Lake Errors



Optimized Hotspot Analysis

Hot Spot Statistics

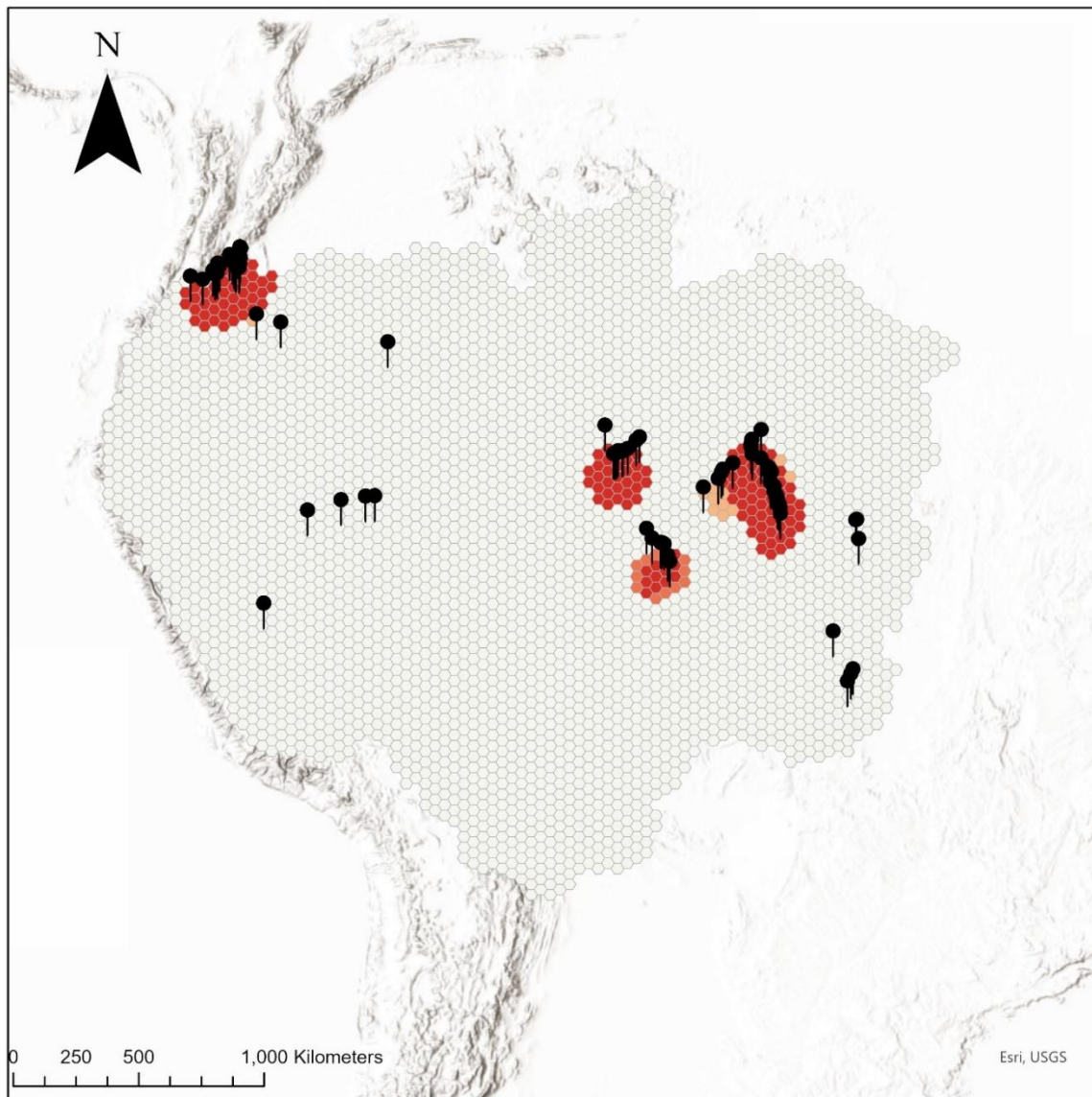
*Getis-Ord Gi**

-  Not Significant
-  Hot Spot with 90% Confidence
-  Hot Spot with 95% Confidence
-  Hot Spot with 99% Confidence

 Oxbow Lakes Error

Figure 5 - Optimized hot spot analysis for 'Oxbow lakes' error

Optimized Hotspot Analysis for 'Missing Centerline' Errors



Optimized Hotspot Analysis

Hot Spot Statistics

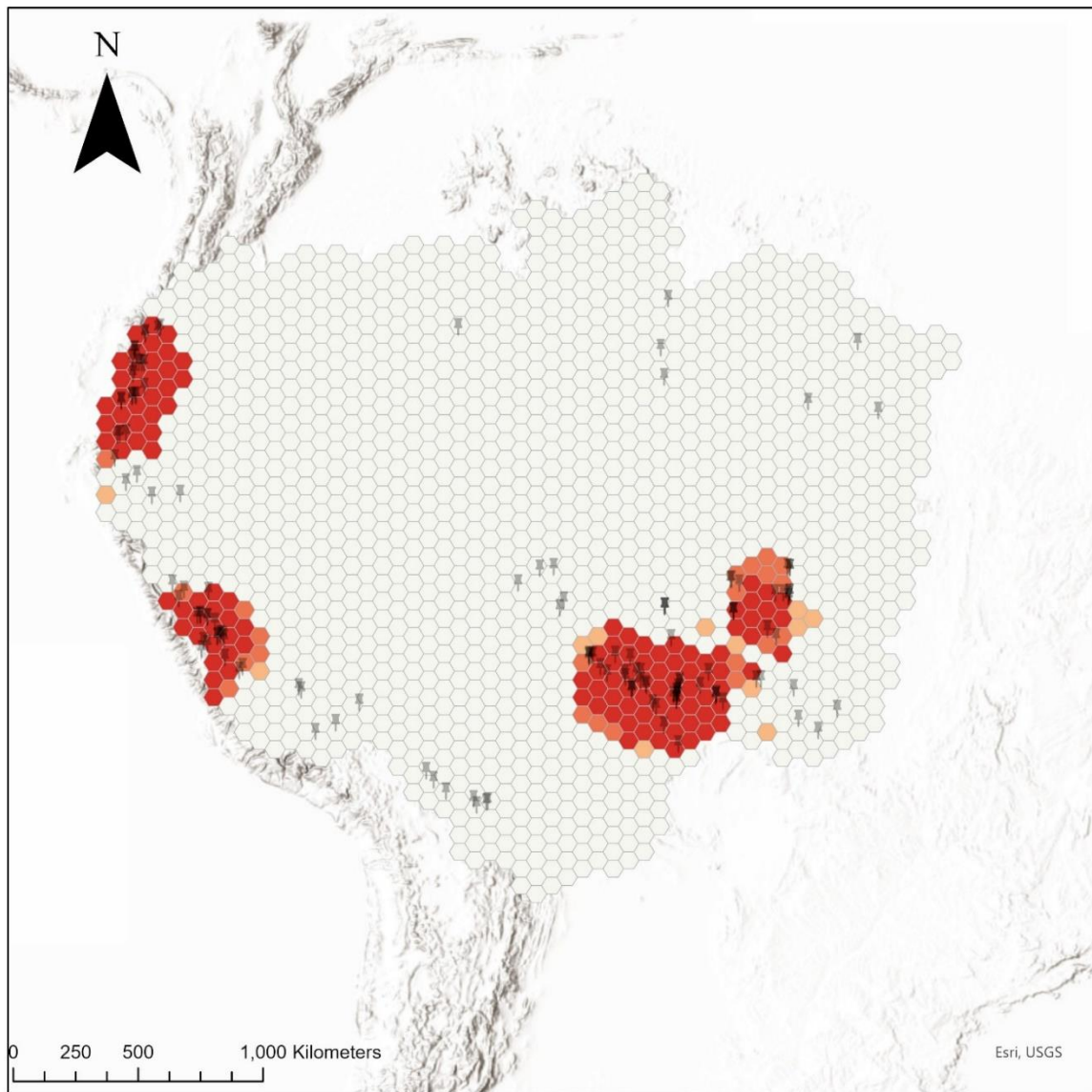
*Getis-Ord Gi**

- Not Significant
- Hot Spot with 90% Confidence
- Hot Spot with 95% Confidence
- Hot Spot with 99% Confidence

Missing Centerline Error

Figure 6 - Optimized hot spot analysis for 'Missing centerline portions' error

Optimized Hotspot Analysis for all Known Dams



Optimized Hotspot Analysis

Hot Spot Statistics

*Getis-Ord Gi**

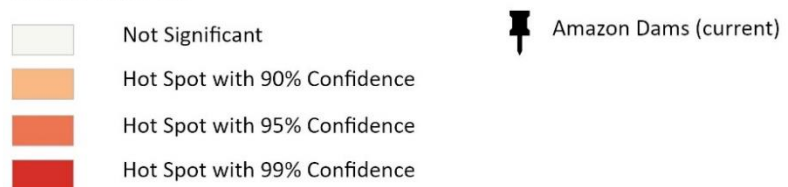


Figure 7 - Optimized hot spot analysis for dam cluster

4.2 Spatial Autocorrelation using Global Morans 1

As previously mentioned, rivers which are under 120 meters in width did not have their centerlines included in the original dataset. Starting from a null hypothesis, confirmation of whether there was a spatial relationship between centerlines ending abruptly and the width of the rivers narrowing was tested. Spatial autocorrelation was run using width as the spatially related parameter. ‘Nearest Neighbor’ conceptualizations were used, along with Euclidian distance methods. The Global Morans 1 statistic used in ArcGIS Pro’s ‘Spatial Autocorrelation’ tool returned a z-score of 2.246774 for abruptly ending centerlines. Global Morans 1 calculates the z-score using the following equation:

$$z = \frac{x - \mu}{\sigma}$$

Eq. 2 - Global Moran’s 1 z-score equation. In the equation, x is the data point, μ is the mean of distribution and σ is the standard deviation of the distribution.

The z-score indicates that the data point is more than two standard deviations above the mean of the distribution. While statistically significant, the dataset is small, which must be taken into consideration. The corresponding p-value calculated was 0.024654, however this was ignored due to the small size of the feature class. The z-score indicated that there was less than a five percent chance that this pattern could have been the result of random chance. A direct correlation was discovered between centerlines ending abruptly and river width.

A Morans 1 index was also calculated. Morans 1 indices range between -1 and 1, with 1 indicating strong evidence of spatial correlation. The index calculated for these parameters was 0.377786, indicating a spatial relationship. This indicates that it was a good experiment choice for Dr. Dente to remove any river width values under 120 meters. This is a complicated issue, however; it should be noted that other factors, such as cloud cover and tree cover become important in analyses using remote sensing techniques.

The missing centerlines issue was also related to river width using the ‘Spatial Autocorrelation’ tool. Once the rivers under 120 meters were masked out, spatial autocorrelation was run using the centerline missing error (error four) as the input dataset and width as the independent variable. A z-score of 7.906170 was returned, indicating a less than one percent chance that this pattern

occurred by random chance. A Morans I value of 0.453954 was also calculated for these parameters, indicating a spatial relationship.

Stream Orders (Strahler) were also correlated to the missing centerline dataset. With stream order being the independent variable, a z-score of 2.995589 was returned, as well as a Morans I value of 0.016274. While this relationship was not as significant as the Moran's value, it still shows evidence of being spatially autocorrelated. Stream order is defined as a value between one through eleven. When two streams of the same order combine, the outlet increases by one order of magnitude. For example, if two order-1 streams combine, the resulting outlet is an order-two stream. A plot of the count of each stream order type has been supplied below (Fig.10). Steam orders of three, four and five combined made up around 87% of all streams in the entire basin.

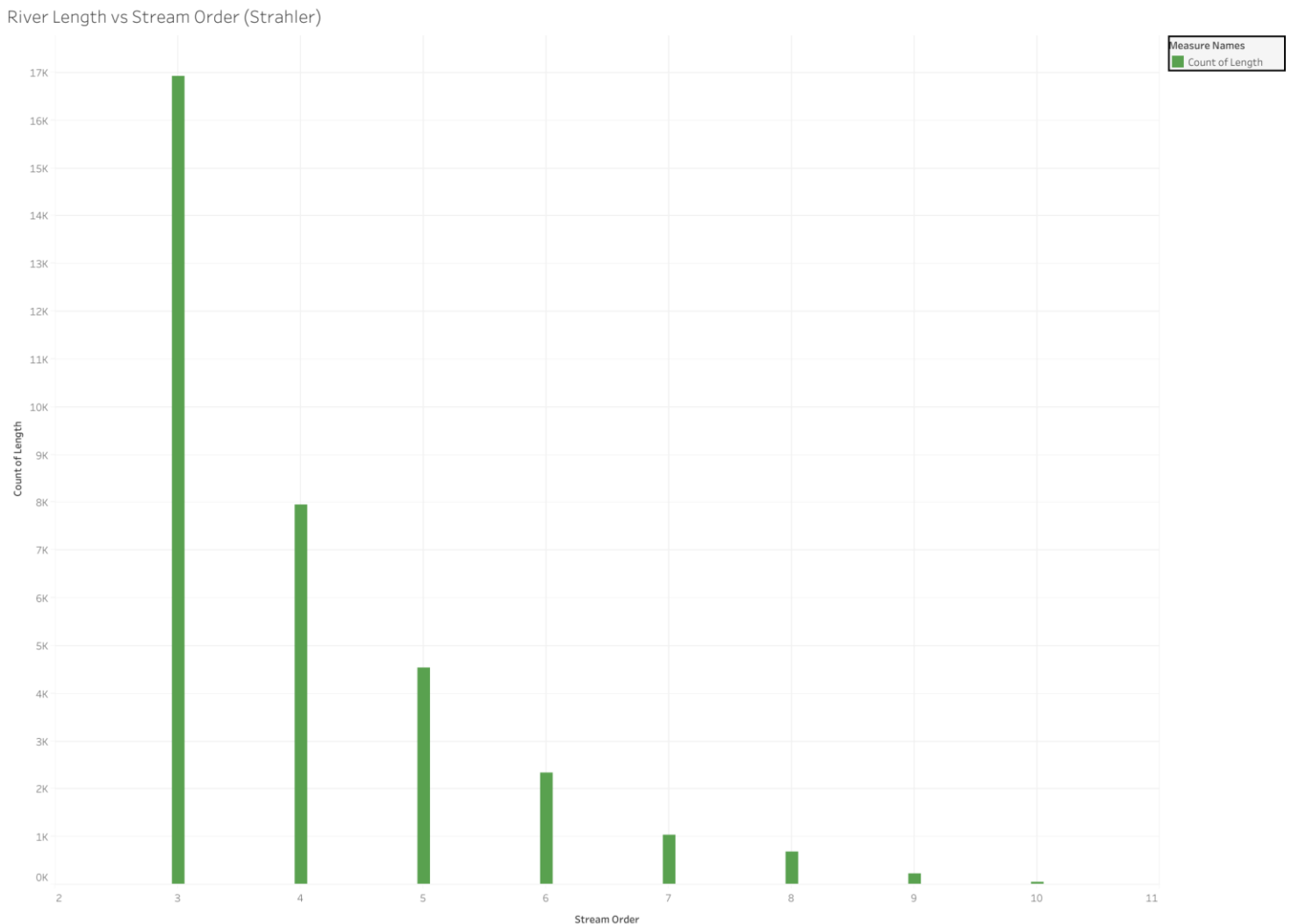


Figure 8 - Stream order count vs length

This is a strong indicator that smaller rivers and streams account for a large portion of the missing centerline errors.

4.3 Hierarchical Density Based Spatial Clustering

HDBSCAN was run on all errors (see Fig. 7). The clustering algorithm first computed the local density of each dataset feature. Local density was defined as the number of other features within a certain radius of the feature. This radius was automatically calculated by ArcGIS Pro. Starting with random features, the algorithm recursively expanded each cluster to include all features within the previously calculated radius. Reachability distance, which is the distance used to determine whether a feature is part of the current cluster, was calculated based on the local density of the feature, using the defined eight-feature parameter. Colors were assigned to each cluster, indicating how many clusters were in each given grouping. The points which were labelled as ‘-1’ were considered by the algorithm to be ‘noise’ and not connected to any cluster. Subsequent listings in the legend indicate clusters of different densities. Lower numbers indicate lower numbers of clusters and higher numbers indicate the opposite. Clustering was not based on the issue ID; however, the Kriging surface gave an idea of which areas have the highest levels of each kind of error. The highest density areas corresponded visually with the outputs from early ‘Optimized Hot Spot’ analysis. A dendrogram has been created in R to visualize the clustering hierarchy. A dendrogram is a ‘tree-like’ plot which shows, in this case, a correlation between the distance and the hierarchy of cluster values. Closer distances tend to correlate with higher clustering values. This is a helpful visualization for spatial correlation. The first law of geography states that “everything is related to everything else, but near things are more related than distant things.” This dendrogram displays this law in practice: the closer phenomena tend to cluster at a higher rate than the father ones.

HDBSCAN Clustering

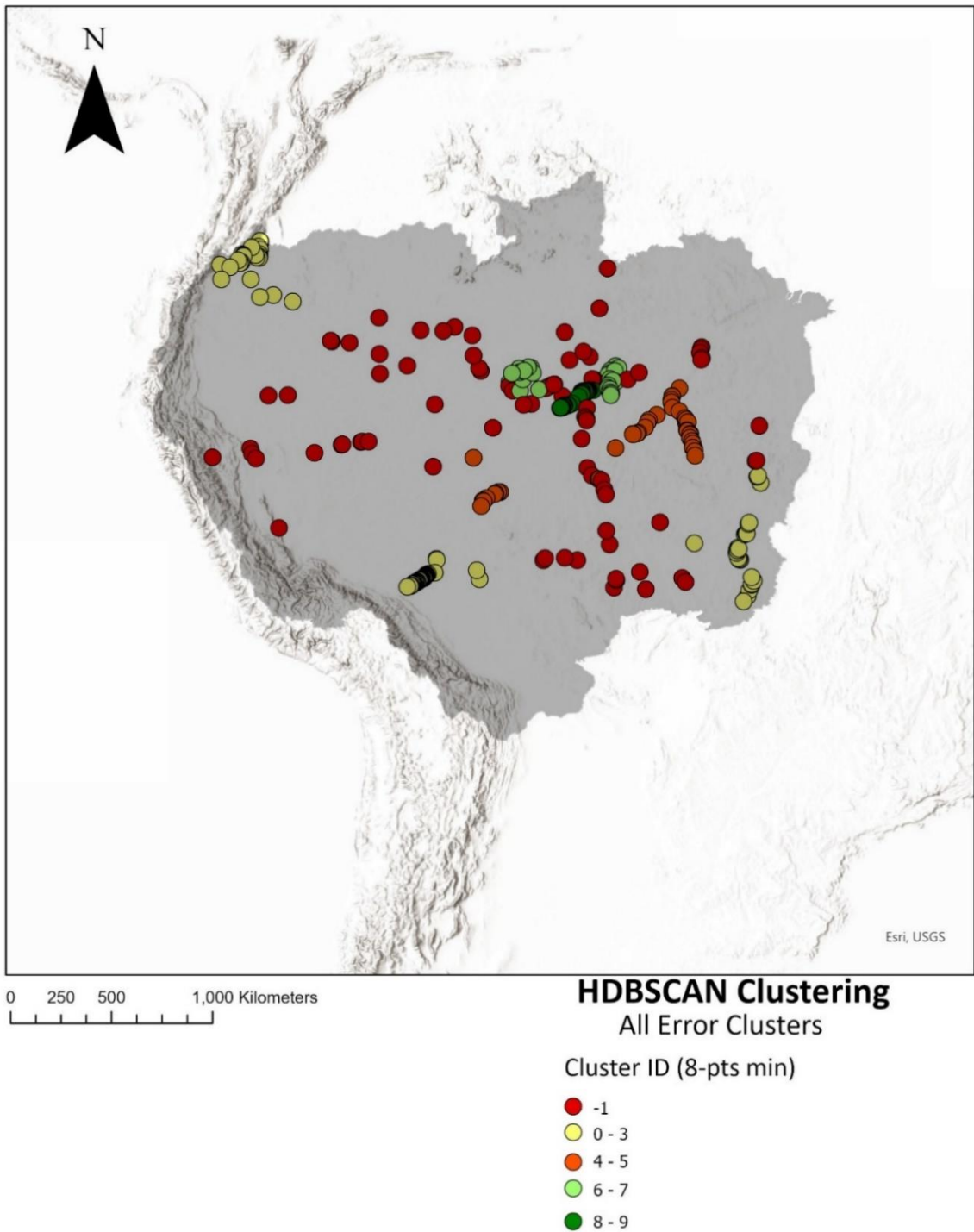
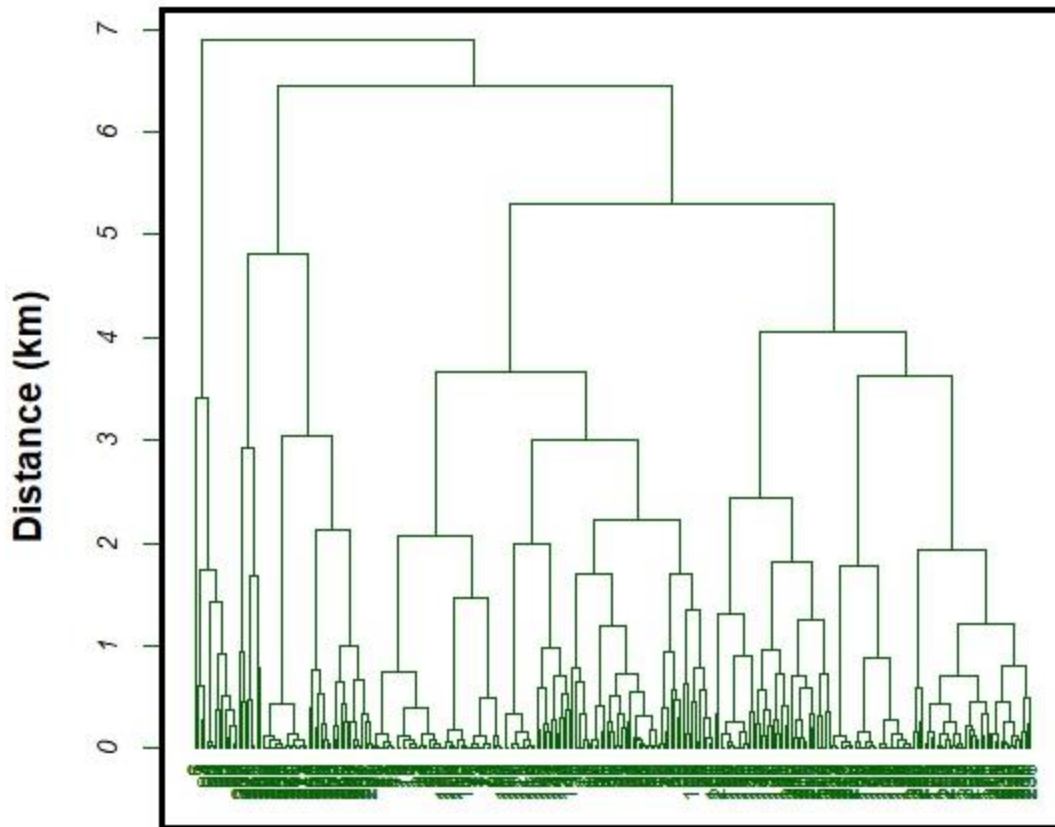


Figure 9 - HDBSCAN Clustering for all errors combined. ID -1 indicates a 'noise' location which is not grouped with any particular cluster. Different cluster colors represent the number of individual points within a meaningful cluster with respect to distance.

Dendrogram for Amazon Basin Error Clusters



Hierarchical Clustering

Figure 10 - Hierarchical clustering dendrogram of the HDBSCAN clusters for all errors. A dendrogram is a 'tree-like' plot which shows, in this case, a correlation between the distance and the hierarchy of cluster values. Closer distances tend to correlate with higher clustering values. This plot indicates cluster amounts with respect to distance between each respective object being evaluated.

4.4 Empirical Bayesian Kriging

EBK was used to create statistical prediction and a corresponding error surface for all errors combined. ArcGIS Pro could not understand that the independent variable was a whole number integer, and so it created a scale based on these integers. The error surface was useful in determining which regions have predicted each error in the strongest way. A power-fit semi variogram was used for this prediction surface. Each error corresponded closely to where the measured ones were in the dataset. The location of reservoir errors may have been skewed when

looking at the dam feature class, but there is a reasons for that. Many of the dams included in the dams dataset are located in the western mountains on headwater streams. Many of these smaller headwater streams were masked out because they were under 120 meters in width. The prediction surface has five value ranges, corresponding to the five error classes. These are 1-1.8; 1.8-2.6; 2.6-3.4; 3.4-4.2; and 4.2-5, respectively. The means of these ranges were calculated and used as the basis for which errors should be most prevalent. The means of the error surface were then calculated and subtracted and added to the means on the predictive surface. Whichever whole number it came closer was chosen as the candidate. After taking the means of the five EBK outputs and subtracting the means of the corresponding error surface values, each surface matched more closely to the error ID. ArcGIS's kriging method and equations are proprietary information, however; general kriging equations look like this:

$$\hat{Z}(S_0) = \sum_{i=1}^N \lambda_i Z(S_i)$$

Eq. 3 - General kriging equation. In this equation, $\hat{Z}(S_0)$ is the measured value at the i th location, λ_i is an unknown weight for the measured value at the i th value location, S_0 is the prediction location and N is the number of measurements.

EBK includes Bayesian methods in their prediction models which includes the idea that prior knowledge can be included in the prediction process. In this case, prior knowledge about the location of existing errors was included in the predicted outcome surface. locations of existing errors. This is most likely due to the use of EBK, as opposed to traditional kriging.

Kriging Surface Output for all Errors Combined

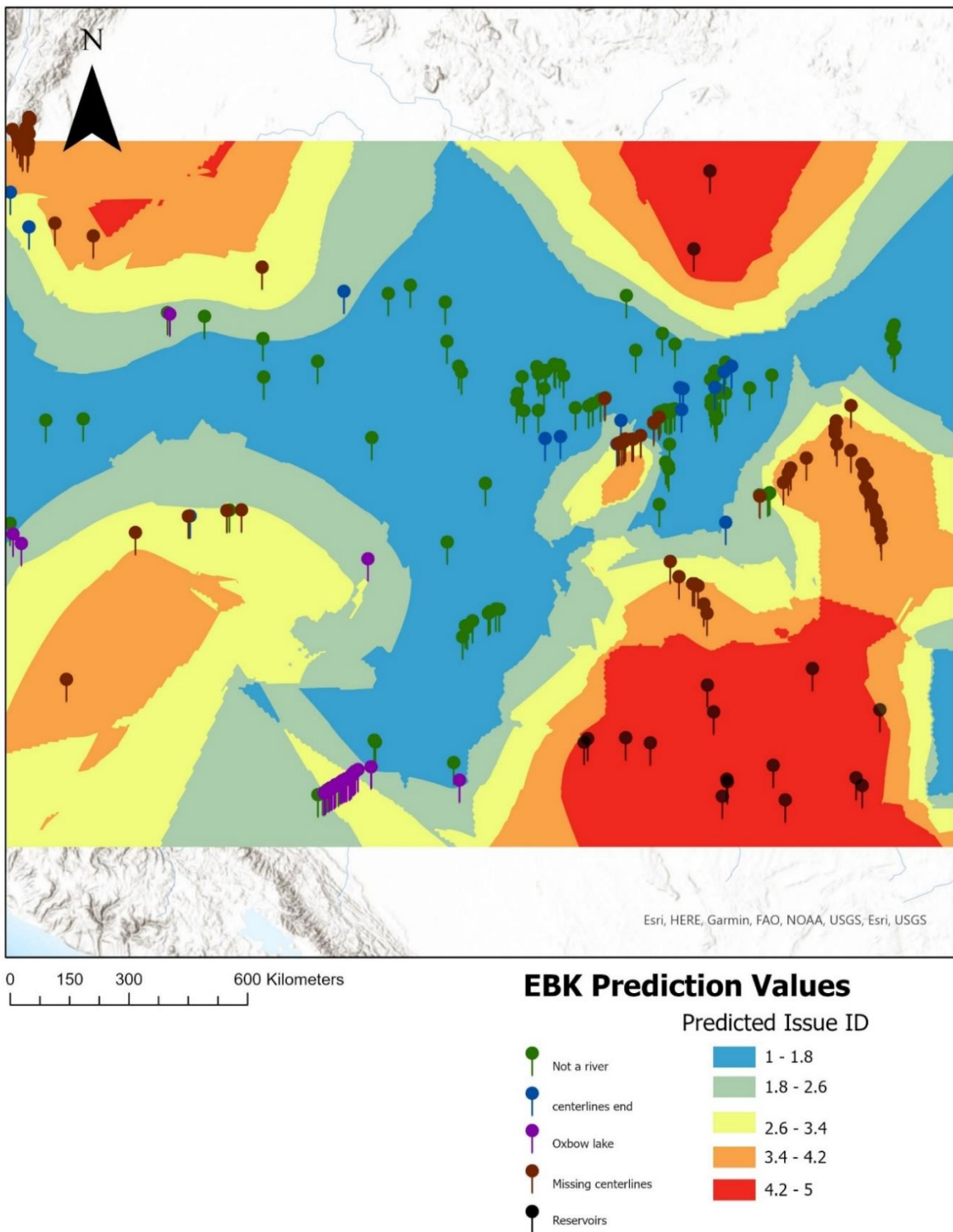


Figure 11 - Empirical Bayesian Kriging surface based on error ID. This surface shows the predicted locations where each error type is suspected to be most common and where new ones could most likely be located.

Kriging Error Surface Output for all Errors Combined

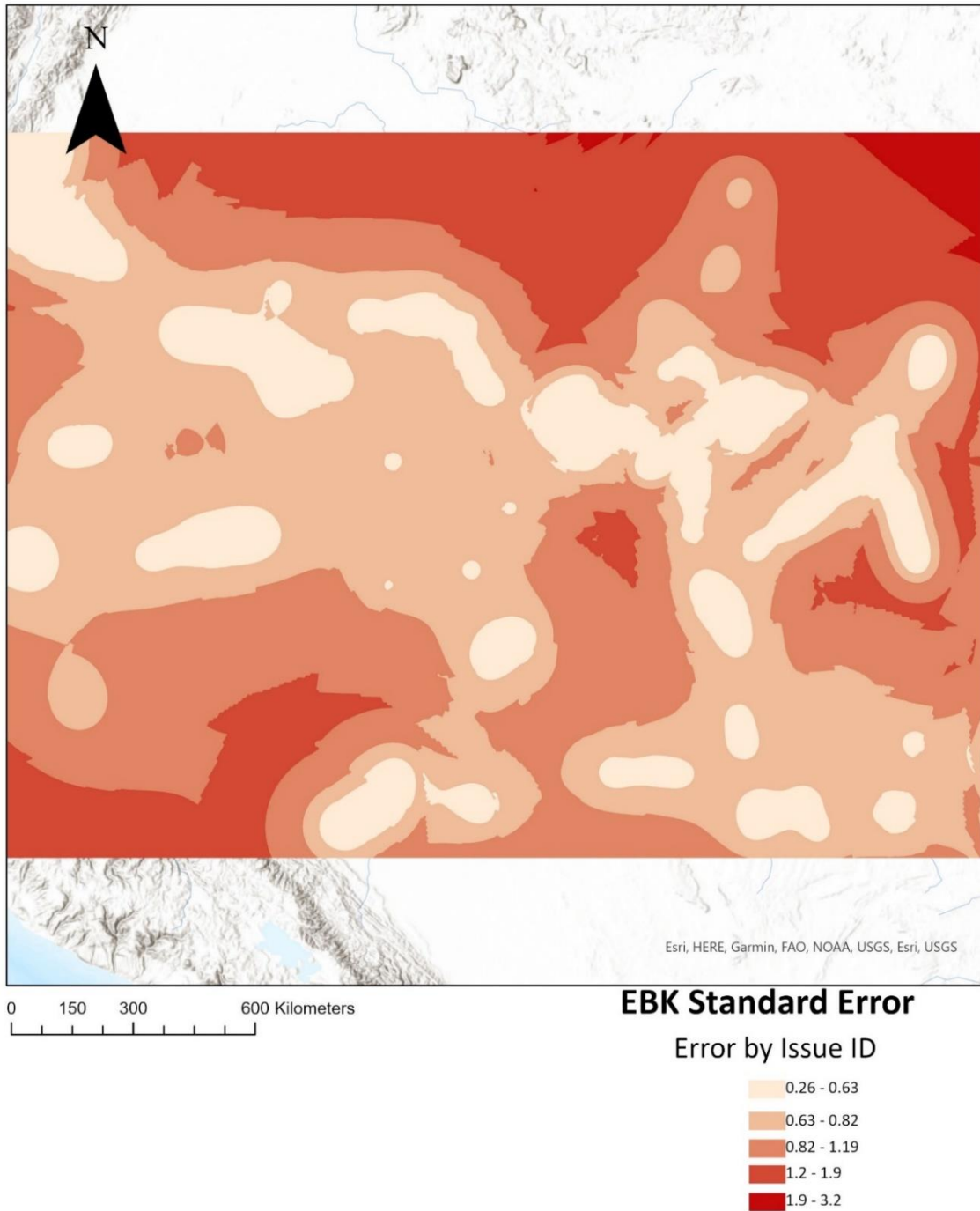


Figure 12 - Empirical Bayesian Kriging error surface based on error ID. This indicates how much standard error is associated with each prediction for each error type spatially. The legend corresponds one-to-one with the legend on the kriging map above.

5. Conclusion

It was determined that there was a strong correlation between river width and errors two and four (centerlines end and missing centerline portion). Z-score indicated that it was very unlikely for these patterns to be random. Additionally, stream order was spatially related to the missing centerline dataset as well; the z-score indicated less than a five percent chance of occurring randomly. Since stream orders 3-5 comprise 87% of all streams in the amazon basin, it was determined that stream width (and overall size) was directly correlated with missing centerline errors. There were also three hot spots for missing centerline errors detected in the center of the basin. This did not necessarily correspond to the area which was originally hypothesized to contain most of these errors. One hot spot was detected in the northwest region of the basin in the areas with steeper stream slopes. The kriging surface predicted that these kinds of errors would be more prevalent in the corners of the maps. Similar to the missing centerline portion errors, prematurely ending centerline errors were determined to have a direct correlation to river width. Significant z-values were derived from using rivers under 120 meters as the independent variable. Many of these errors occurred because a small river would widen to 120 meters or wider for short spans, causing centerlines to be calculated for these small portions. Areas which were determined to be rivers but were not found had hotspots in the center and southeast regions of the map. The kriging prediction aligned with the actual surface, as well as the hypothesized hotspots well. The lakes in the region which were misclassified tend to be concentrated around the hot spot areas. Oxbow lakes have a hotspot in the southern region, as expected. This area contains many braided, meandering rivers. They could not be correlated to the SWORD dataset, since they were no longer a part of the actively flowing river. Reservoir hotspots originated in regions where there were high levels of deforestation and human activity. These regions are mostly in the southern part of the basin. This was expected, since reservoirs are created by dams. The kriging surface correlates heavily with the dam clusters. Understanding why these errors occur and what kinds of regions are most prone to them is crucial for improving this technology for future use. Once problematic areas are identified and classified globally, improvements will be able to be made to automatic aping techniques.

Special thanks to Dr. Elad Dente for allowing me to use his current data, as well as for his guidance.

Appendix A: Source Code

```
#cbind into scaled matrix
HDBSCAN_2<-cbind(scale(cbind(HDBSCAN_all_errors[,1:2],HDBSCAN_all_errors[,5:7])))
#compute distance matrix
d <- dist(HDBSCAN_2)
#hierarchical cluster
hc <- hclust(d, method = "complete")
# Plot the dendrogram with custom parameters
plot(hc,
      main = "Dendrogram for Amazon Basin Error Clusters",
      col = "dark green",
      hang = -1,
      labels = HDBSCAN_all_errors$OBJECTID,
      xlab = "Hierarchical Clustering",
      ylab = "Distance (km)",
      font.lab = 2,
      cex.lab = 1.2,
      cex.axis = 0.8,
      cex.main = 1.5,
      font.main = 4,
      sub = "",
      hang.leaf = TRUE,
      cex = 0.7,
      main.col = "black",
      sub.col = "black",
      font.axis = 3,
      font.labels = 2,
      col.axis = "black",
      col.main = "black",
      col.sub = "black",
      col.labels = "black"
)
box(lwd = 3)
R- Source code for hierarchical cluster analysis and dendrogram
```

Appendix B: Supporting workflows

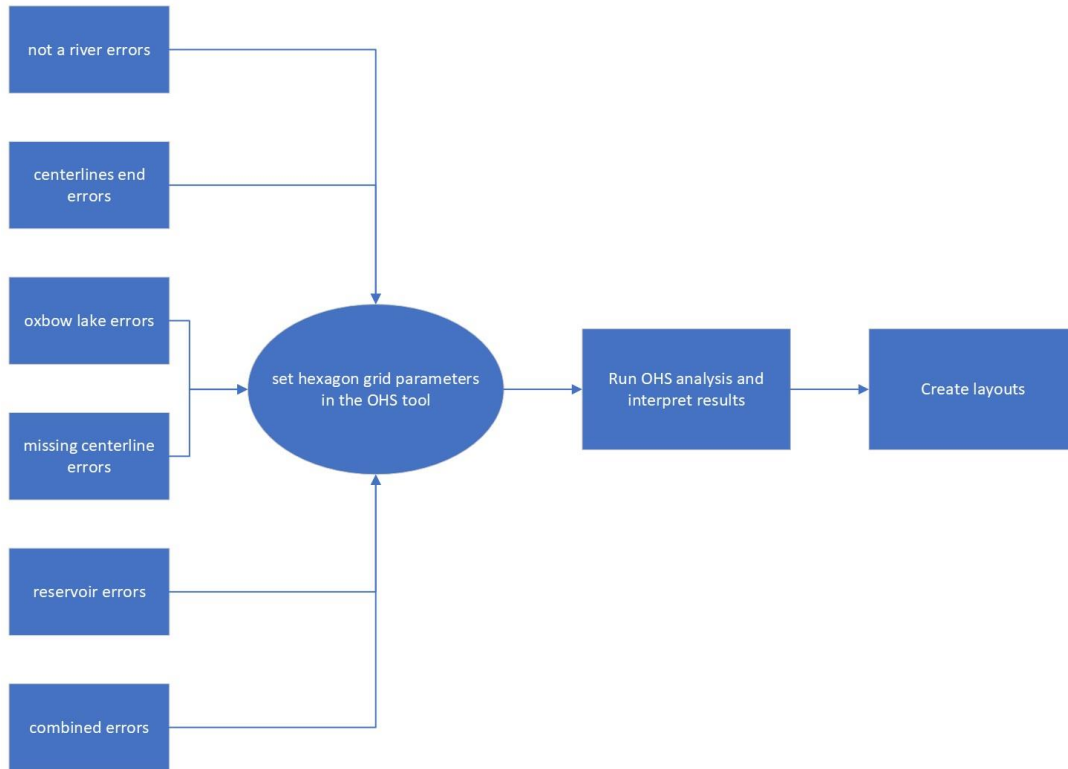


Figure 13 - OHS analysis workflow

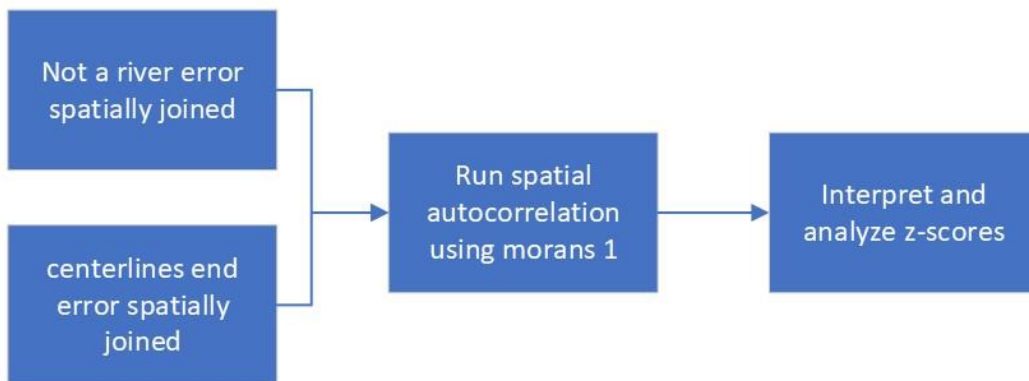


Figure 14 - Spatial autocorrelation workflow

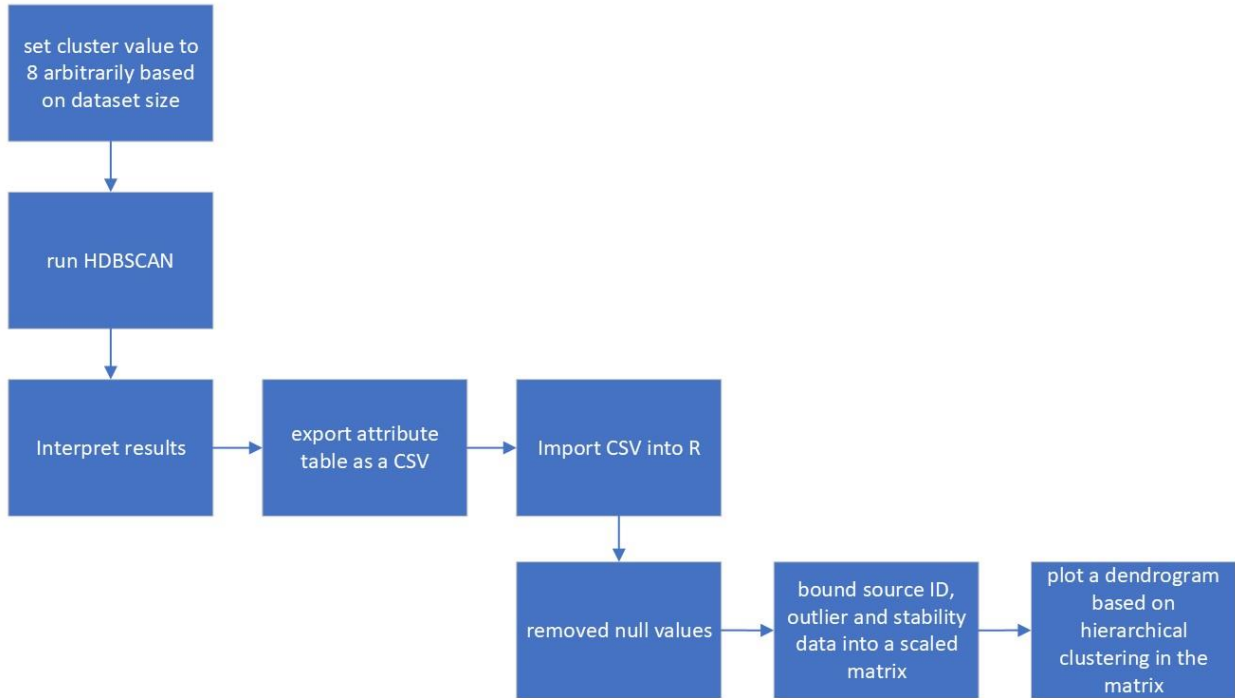


Figure 15 - HDBSCAN workflow

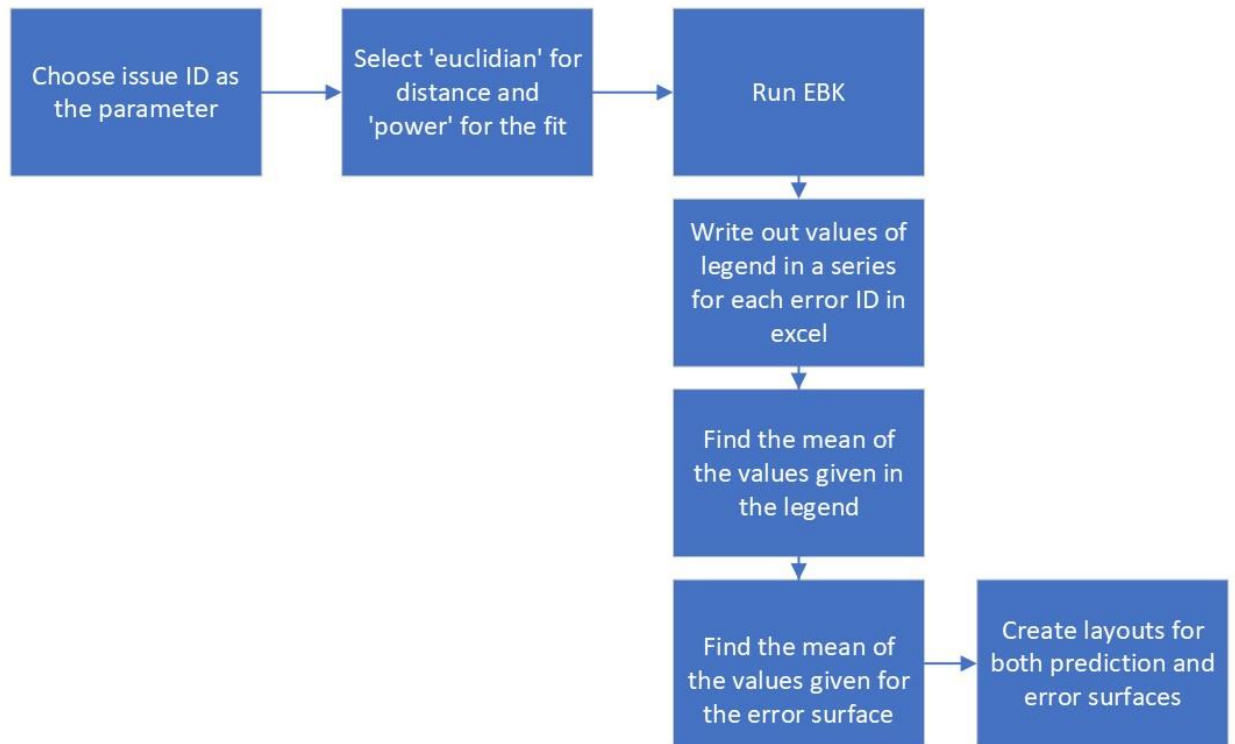


Figure 16 - EBK workflow

Works Cited

- Agarwadkar, A.M., S. Azmi, and A.B. Inamdar. 2013. Understanding grids and effectiveness of hexagonal grid in spatial domain. In Proceedings of the International Conference on Recent Trends in Information Technology and Computer Science (ICRTITCS –2012), 17–18 December 2012, Mumbai, India, 25–27.
- Altenau, E. H., Pavelsky, T. M., Durand, M. T., Yang, X., Frasson, R. P. de M., & Bendezu, L. (2021). The Surface Water and Ocean Topography (SWOT) Mission River Database (SWORD): A Global River Network for Satellite Data Products. *Water Resources Research*, 57(7). <https://doi.org/10.1029/2021wr030054>
- Elizabeth H. Altenau, Tamlin M. Pavelsky, Michael T. Durand, Xiao Yang, Renato P. d. M. Frasson, & Liam Bendezu. (2022). SWOT River Database (SWORD) (Version v14) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7410433>
- Giustini, F., Ciotoli, G., Rinaldini, A., Ruggiero, L., & Voltaggio, M. (2019). Mapping the geogenic radon potential and radon risk by using Empirical Bayesian Kriging regression: A case study from a volcanic area of central Italy. *Science of the Total Environment*, 661, 449–464. <https://doi.org/10.1016/j.scitotenv.2019.01.146>
- Grubestic, Tony & Murray, Alan. (2001). Detecting Hot Spots Using Cluster Analysis and GIS.
- Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., & Townshend, J. R. G. (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160), 850–853. <https://doi.org/10.1126/science.1244693>
- Horton, A. J., Constantine, J. A., Hales, T. C., Goossens, B., Bruford, M. W., & Lazarus, E. D. (2017). Modification of river meandering by tropical deforestation. *Geology*, 45(6), 511–

514. <https://doi.org/10.1130/g38740.1>

How Density-based Clustering works—ArcGIS Pro | Documentation. (n.d.). Pro.arcgis.com.

<https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm>

Krivoruchko, K., & Gribov, A. (2019). Evaluation of empirical Bayesian kriging. *Spatial Statistics*, 32, 100368. <https://doi.org/10.1016/j.spasta.2019.100368>

Mayorga, E., M.G. Logsdon, M.V.R. Ballester, and J.E. Richey. 2012. LBA-ECO CD-06

Amazon River Basin Land and Stream Drainage Direction Maps. Data set. Available online [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A.

<http://dx.doi.org/10.3334/ORNLDAAC/1086>

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633), 418–422.

<https://doi.org/10.1038/nature20584>

Roberts, S. A., Hall, G. B., & Calamai, P. H. (2000). Analysing forest fragmentation using spatial autocorrelation, graphs and GIS. *International Journal of Geographical Information Science*, 14(2), 185–204. <https://doi.org/10.1080/136588100240912>

Zerbe, K., Polit, C., McClain, S., & Cook, T. (2022). Optimized Hot Spot and Directional Distribution Analyses Characterize the Spatiotemporal Variation of Large Wildfires in Washington, USA, 1970–2020. *International Journal of Disaster Risk Science*, 13(1), 139–150. <https://doi.org/10.1007/s13753-022-00396-4>