An Automated Method for the Detection of Topographic Patterns at Tectonic Boundaries

Bobak Karimi
Department of Geology
Colorado College
Colorado Springs, Colorado, USA
Email: bobby.karimi@coloradocollege.edu

Hassan A. Karimi
School of Information Sciences
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
Email: hkarimi@pitt.edu

***Abstract*—Detection of spatial and contextual patterns is of great importance to geoscientists interested in understanding and analyzing tectonic boundaries. To date, geoscientists have developed mostly manual detection methods and only recently has interest in the development of automated methods grown with the availability of high-resolution satellite data and the advancement of technologies such as Geographic Information Systems (GIS). Geoscientists are examining different approaches to automate the manual detection method of tectonically related phenomena, but considering the complexity, time-consumption, and assumptions usually made in the manual method, new automated detection techniques are anticipated to surface soon and will vary in implementation, accuracy, and time performance. In this paper, we present a Digital Elevation Model (DEM) based automated method for detection of spatial and contextual topographic patterns at tectonic boundaries. Our automated method was experimented and compared against recent existing methods with the same objective and the manual method, which is considered as the baseline. The results show that our automated method produces more accurate results than the existing methods.**

Keywords - Automated pattern detection; cluster analysis; lineaments; tectonics.

I. INTRODUCTION

Context plays a major role in detecting patterns and can help improve the accuracy of the automated detection methods. In particular, spatial context is necessary and imperative in detection of patterns in natural phenomena. In this paper, we discuss a new automated detection method in the context of lineaments (such as faults). We chose this context, as a representative spatial context, primarily for the reason that it is complex and involves several steps, including image processing for pixel extraction from raster datasets, conversion of extracted pixels to vector lines, and detection of line clusters based on the context.

Tectonic stresses acting upon a region can create deformation structures, such as folds, faults, and fractures. These structures act as pathways for weathering and erosion, influencing topographic pattern development in a region. Selecting lines (or lineaments) along changes in topography (ridges, valleys, etc.) is a common method in highlighting patterns of geologic structures. The scale of the lineaments often reflects the most prominent types of geologic structures [1]–[3]. Detection and selection of lineament patterns is most accurately conducted manually; however, automated methods are improving and their results are converging on the accuracy of the manually selected data. These automated methods still are prone to error, and validation of the dataset is crucial. Since large, tectonically sourced lineaments are not just expressions of a specific feature, but also a manifestation of lithospheric paleostress fields [4]–[7], errors in the orientation of lineaments can lead to incorrect interpretations of the geologic stress patterns through time [4].

The most common methods of automatically extracting lineament data is from DEMs [8], [9], or from some surface derived from a DEM [10]–[12]. Derived surfaces are created to better enhance distinct topographic changes (‘edges’), and the derived surface most frequently used are hillshades [4], [11], [12]. The popularity of utilizing hillshades to select lineaments is based on how well it highlights topographic changes; however, since a hillshade is based on azimuthal direction and vertical angle of a ‘sun’, a single image will prominently highlight features perpendicular to the azimuthal orientation selected to make the hillshade [13]. Features not ideally oriented will be harder to select as the difference in Digital Number (DN) value across the ridge will be close to negligible. To overcome this hurdle, several hillshades with different azimuthal sun orientations can be created from the original DEM [13], [14]. Once edges have been enhanced, edge linking methods are used for automated line extraction [4], or modules such as LINE in PCI Geomatica.

The line datasets selected from the various hillshades sometimes highlight the same topographic features, and the resulting final dataset has clusters of lines representing a single feature. Manually picking lineaments from the hillshades avoids this data clustering [13]; however, these clusters are unavoidable in automatically selected lineament data based on a multiple hillshade approach, requiring a method to de-cluster or assess the data. Assessment of automatically picked lineaments has most commonly been done subjectively as a visual assessment [12], [15]. There have been several objective approaches suggested: [4] used a hierarchical clustering of different datasets based on count and statistics of orientation and length of lineaments, similarly, [16] computed statistics of count and length of lineaments to compare different datasets, [8] implemented a confusion matrix approach with the distance between lineaments, and [9] used calculated reference point data to correlate with ground truth datasets as a comparison metric. While these objective methods deploy specific metrics to assess or de-cluster data,

 ISBN: 978-1-61208-534-0

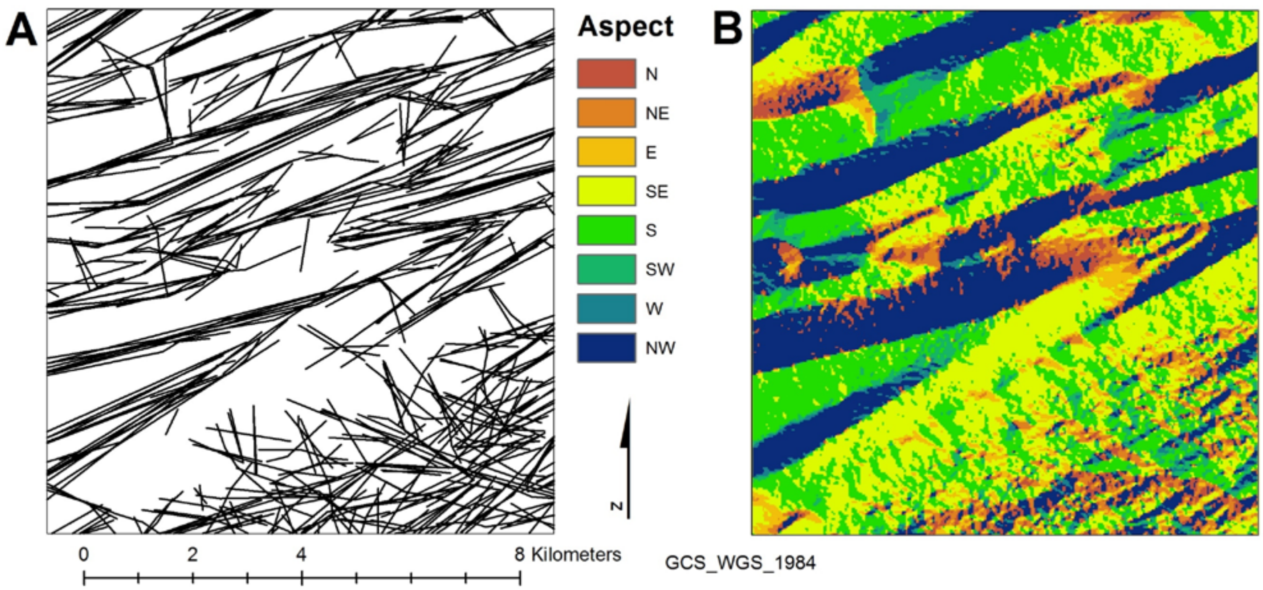

Figure 1. (a) The final dataset of all lines extracted from the multiple hillshades in PCI Geomatica with noise reduction, and (b) Aspect of the DEM.

none reference the original dataset, the DEM, to evaluate lines within a cluster as the most representative of the topographic feature they are meant to represent. Without consideration of what the lines represent, any de-clustering method or assessment allows for misoriented linear evaluations of topographic features, leading to misinterpretations of geologic stresses.

The objective of this paper is to present a new method in assessing the different datasets that result from a multiple hillshade (MH) lineament selection approach, referencing the original DEM dataset as an objective assessment of lines. Our results are to be compared to the leading DEM based lineament method by [4] and a manual method, which we consider to be the baseline. The manual method, despite being very time consuming and in some parts involving subjective assumptions, is currently the only known method that produces highly accurate results. This, coupled with the observation that the geoscience community has spent many years refining and improving the manual method, would make it suitable to be used as a baseline for comparing the accuracy of automated methods, such as the one discussed in this paper.

The rest of this paper is structured as follows. In section 2, we discuss the study area and the digital elevation model we use. Section 3 outlines our methods: automated selection of lines, derivation of metric to validate lines, and cluster analysis. Our results are presented in section 4 and discussed in section 5. Concluding remarks and future work are found in section 6.

II. Study Area and Data

The eastern margin of North America has been subjected to several mountain building (orogenic) events over the last 500 million years, the final event constructing the Appalachian Mountains. For the purposes of this study, we locate our region of interest to an approximately 9km x 9km region in central Pennsylvania within the Valley and Ridge province of the Appalachian Mountains. This area is dominated by folded beds whose preferential weathering and erosion dominate the topographic development in the area. Ridges and valleys within this particular area trend NE-SW at an azimuth of 67°. Our DEM is a 1-arc second (~30m) resolution elevation model from the Shuttle Radar Topography Mission (SRTM) V2. Within the DEM, there are two distinct topographic expressions. In the northern and western portions, there is a distinct topographic representation of the valleys and ridges associated with the province, with high topographic relief. The southeastern portion has no pronounced ridge system and the topographic relief in this region is low. It is likely that the structures expressed by topography in this region are joints (fractures).

III. Methods

Our new method discussed in this paper is similar to the Multi-Hillshade Hierarchical Clustering (MHHC) method presented in [4]. The selection of clusters follows similar steps, but our methods vary in how we process those clusters. Our method is composed of three parts: 1) automatic selection of lines in PCI Geomatica, 2) derivation of a metric to validate the best oriented lines, and 3) the cluster analysis.

A. *Automatic Selection of Lines*

The automatic selection of lines is composed of four steps: 1) creation or acquisition of a DEM, 2) derivation of hillshades from the DEM at various illumination azimuths, 3) line extraction based on edge detection, and 4) reduction of noise. The DEM can be created from vector or LIDAR data, or acquired as a subset or mosaic of existing DEMs. In our case, we utilized a subset of a DEM with coverage in central Pennsylvania. From the DEM, we derived eight hillshades at 45° illumination orientations starting with 0° (north) and ending with 315° (northwest). We selected these eight orientations to best highlight topographic changes

 ISBN: 978-1-61208-534-0

(ridges/valleys) that may otherwise not be well-highlighted given only a single illumination orientation. For example, a ridge oriented east-west will best be highlighted with a north-south illumination and not an east-west orientation. Each image was then imported into PCI Geomatica, calling on the LINE module to extract lineaments. The LINE algorithm is comprised of three steps. The first is the edge detection operator (Canny edge detector) followed by thresholding to produce a binary edge raster [17]. This image is then processed by many substeps to extract the vector lines [4], [17]. Further, and more detailed, description of the workflow for the LINE module can be found in [17]. The LINE module requires several parameters to be input, and these parameters can impact the count, length, and spatial accuracy of the selected lines [4]. Parameter selection was based on several trials and visual assessment of the output. The parameters we selected are provided in Table 1. The values in Table 1 are expressed in pixels (px) as these are the values of inputs used by the software. A vector shapefile of automatically selected lineaments is output for each of the eight hillshade images. These shapefiles were merged into a single dataset, and each line was split at vertices. Splitting the lines increases the number of lines and the dataset size, but it also allows for interpretation of multiple structures influencing a single topographic feature. Azimuthal orientation and length of each line was calculated and added as a field to the dataset.

Noise reduction was performed using a raster approach outlined in [4]. The merged dataset was converted to a raster image using the line density tool in the computer program, ArcMap [18]. In the output raster, considering a relatively low value for search radius, clusters of lines are depicted as regions with high DN values, while solitary lines have lower DN values. Zonal statistics of the line density output raster were calculated within a 60m buffer around each line and the mean DN value was appended to the line dataset. Lines associated with low DN values were deleted from the dataset. It must be noted that, while this noise reduction decreases the total number of lines, it also may remove small, but structurally relevant data. At this stage of the research, we continued with the noise reduction, as it is what was employed by [4] in their method. As our work continues, we will need to make considerations of the validity of noise reduction in the context of geologic structural and field data. The final dataset is shown in Fig. 1A. Clusters of lines can clearly be seen that follow ridge/valley profiles.

TABLE I. TABLE OF PARAMETERS FOR LINE SELECTION IN PCI GEOMATICA.

Parameter	Value Used
Filter Radius	10 px
Edge Gradient Threshold	30 px
Curve Length Threshold	30 px
Line Fitting Error Threshold	9 px
Angular Difference Threshold	30°
Linking Distance Threshold	20 px

B. *Derivation of Metric to Validate Lines*

In deriving a metric to validate lines, special attention must be paid to what these line features represent, which, based on the MH approach, are changes in topography. Not all lines within a cluster are true representations of the topographic feature they are meant to highlight. The most accurately oriented line will have slopes oriented differently on both the right and left side of the line, as that line represents some valley or ridge. This will not hold true for lines inaccurately oriented, as the same slope orientation can exist on both sides of the line. To implement this idea and develop a metric which we refer to as the *inflection value*, we first derive an aspect image from our DEM (see Fig. 1B). This provides us with the azimuthal orientation each slope is facing within the DEM. The aspect image is converted to a point shapefile. We then create 60m left and right buffers around each line. Unfortunately, the zonal statistics tool does not take into consideration circular statistics, so we calculated the sine and cosine for the azimuth at each point and developed our own zonal statistics tool specifically for circular data. This tool uses the aspect point shapefile and the left and right buffers as input and calculates the mean of the sine and cosine within each buffer, and converts that value back to azimuthal notation. The difference between the mean aspect values in the left and right buffers of each line gives us the *inflection value*. High *inflection values* represent lines that more accurately represent the feature they are meant to highlight, since the slopes on either side of a ridge or valley should face opposite directions.

C. *Cluster Analysis*

The following workflow, adapted from [4], is applied to the line dataset to reduce clusters to one linear feature:

1) Choose the longest line in the main dataset.
2) Make a buffer around the chosen line.
3) Select all lines completely within that buffer.
4) Select lines with azimuth within 20° from the line selected in step 1.
5) Select the line with the largest inflection value, save it to a new shapefile – the final dataset - and delete all selected lines in the main dataset.
6) Repeat from step 1 until no lines remain in the main dataset.

We compare our cluster analysis with that suggested in [4], which follows a similar workflow up until step 5:

5) If the selection contains more than 4 lines, continue to step 6, otherwise save the originally selected line to a new shapefile – the final dataset – and continue to step 8.
6) Create a buffer around selected lines (=cluster) with the following attributes: count of selected lines, average length, and average azimuth.
7) Create a new line using the average length and azimuth in step 6, and save it to the final dataset shapefile.
8) Delete all selected lines from the main dataset.

9) Repeat from step 1 until no lines remain in the main dataset.

In deploying the method outlined in [4], we had to explore and make some assumptions as to the size of buffers. The buffer size in step 2 is determined by processing the dataset using the two methods using 150m, 200m, 250m, and 300m buffers. The results are visually assessed to evaluate the buffer size that best de-clusters data; where clusters are reduced to a single line and not an excess of lines are deleted. In step 6 for [4], we utilized a buffer of 60m, which was large enough to allow all buffered lines to intersect one another. Beyond these assumptions, we maintained the exact process as described in [4] to better compare the two methods and assess sources of difference. For both methods, the algorithms were written in Python in ArcGIS using the ArcPy library.

IV. Results

The results of the method by [4] using 150m, 200m, 250m, and 300m buffers at step 2 are shown in Fig. 2. Similarly, Fig. 3 shows the results using our new method. The lines in these images represent final de-clustered line datasets within the region of interest using different buffer sizes in step 2.

Both sets of results were compared to a more accurate lineament dataset manually selected using hillshades and the original DEM. Through visual assessment, we are able to identify that a 300m buffer resulted in too few lines in the final dataset in both methods. Lower buffer sizes (150m and 200m) left too many lines representing single features. A buffer size of 250m provided the best results in the case of both methods. Results of both methods compared to the manually selected lines are shown in Fig. 4. In Fig. 4, the top rose plot is of the manual data, middle rose plot is of the data from our method, and the bottom is from the method in [4].

V. Discussion

We calculate a completeness percentage of the resulting output lines from the two methods as compared to a manually selected dataset [19]. This calculation was done by buffering the lines from the two methods with a buffer size of 50m, and extracting the length of manual lines within those buffers. The percentage of the length of lines within the buffers is the completeness percentage. Our method had a 60% completeness compared to 47% that resulted from the method in [4]. Upon a visual assessment of the lines output by the two methods using a 250m buffer size in step 2 to manually selected lines (see Fig. 4), we can note that automatic lineament selection in the northern portion of the dataset was far more successful than what is seen in the southeastern region. Completeness percentages were higher with both methods in the north: 80% for the results of our method, and 68% for the method by [4]. We compared our initial total dataset before the clustering method was applied, and were able to ascertain that the cluster analysis caused the significant loss of data in the southeastern region. This leads us to believe

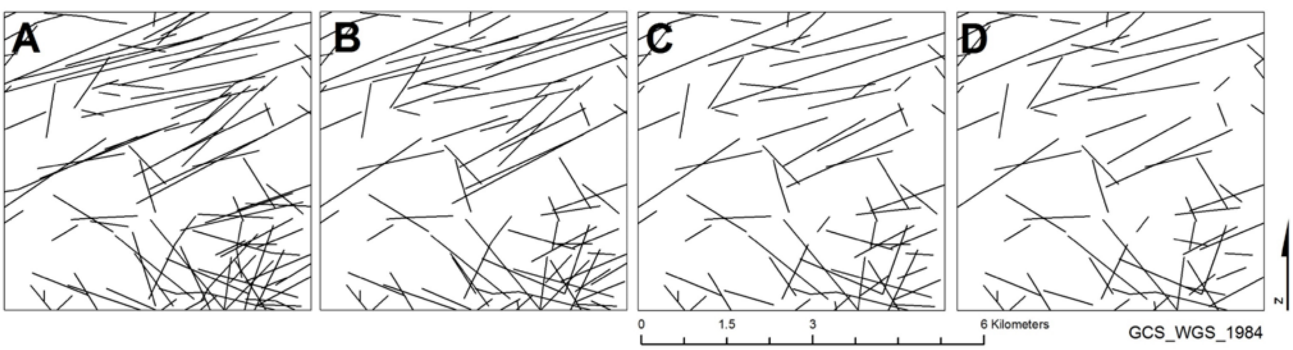

Figure 2. Results of the method in [4] at (a) 150m, (b) 200m, (c) 250m, and (d) 300 m buffers.

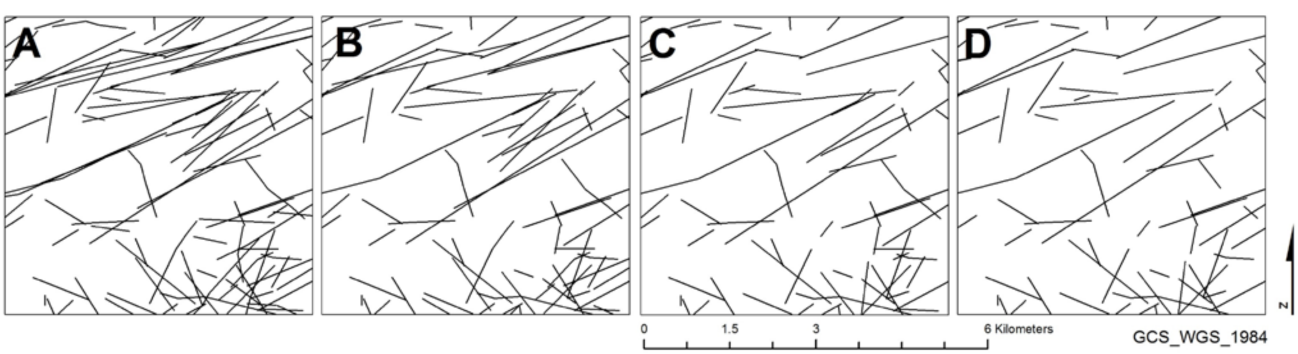

Figure 3. Results of our method at (a) 150m, (b) 200m, (c) 250m, and (d) 300 m buffers.

that subdivisions of the dataset should be made based on a first pass visual assessment of the automatically selected line dataset. These subdivisions can then be processed using different buffer sizes, or even approaches, to produce a more accurate result for that subarea. Since the southeastern portion of the area of interest has been identified as compromised, we make our assessment on the effectiveness of the two methods based on the northern portion.

In the northern region, both methods do not highlight every line, but this happens where the buffer size to select a cluster overlaps lines representing an adjacent feature with similar orientation and size. One potential way to avoid this in the future could be to remove the necessity for buffering at step 2 and only select lines that intersect the longest line. Not all lines within a cluster intersect every other line, so this could lead to additional errors. Additionally, lines in clusters near the ends of another cluster could intersect the longest line as well.

A visual assessment between the three datasets in the northern area suggests that our method more often picks lines that match in orientation with the manual dataset. By creating a metric for each line that references the original dataset, we have provided a new method in differentiating the most representative line in a cluster. Beyond quantitative and visual assessments, rose plots have been created for the three datasets (see Fig. 4). These rose plots represent the frequency of azimuthal orientations of lines in the overall dataset. The manually derived dataset has a clear east-northeast trend that is bimodal (peaks at 60° and 75°) within a range of 30°. This general trend is shared with our method, and the method from [4]; however, the bimodal characteristic with similarly oriented peaks is seen only in the rose plot of our new method.

This combination of quantitative, visual, and data trends assessment leads us to suggest that our new method is better in differentiating datasets (MH lines). Additionally, our new method highlights the importance of referencing the original DEM when validating or assessing clusters of lines. While our

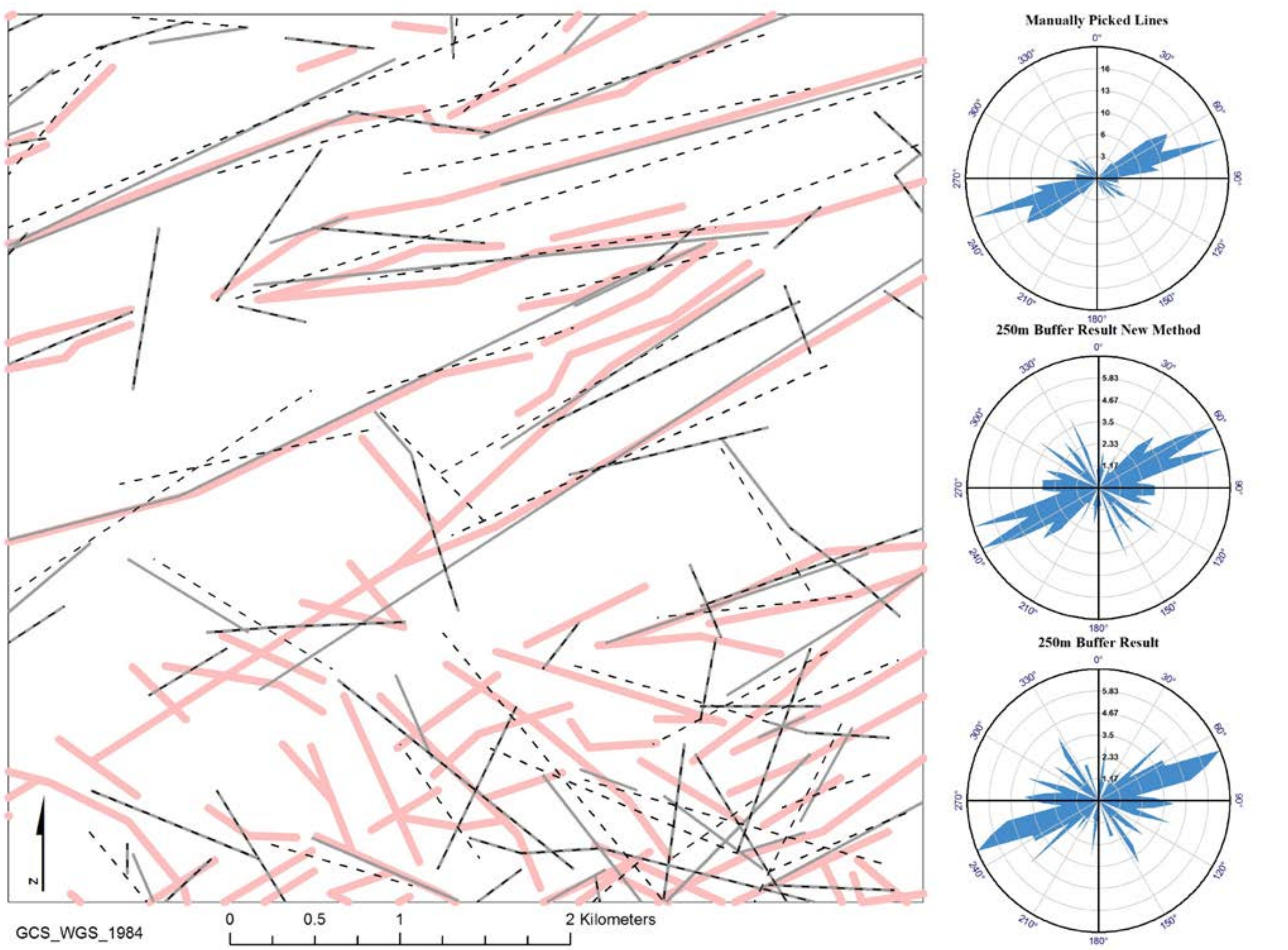

Figure 4. Comparison of the manually picked dataset (red lines) to our new method (solid grey lines) and to the results of the method in [4] (dotted black lines) using a 250m buffer. Rose diagrams are provided for each dataset.

method produces more accurate results, there are still many improvements to be considered, such as avoiding the loss of adjacent data with similar orientations and lengths. Since the algorithms for these improvements are computationally complex, processing large datasets would take an enormous amount of time. Work has to be done on creating a more time-efficient approach. Furthermore, we hope to explore more quantitative and automated methods of parameter selection where parameter selection is based on trial-and-error and subjectivity, such as the input values for the module used in PCI Geomatica (Table 1).

VI. Conclusion and Future Research

We have successfully proven that referencing the original DEM when assessing line data within clusters results in a more accurate representation of features in a region. However, our method requires adjustments to take into consideration distances between clusters, and how regions dense with data (southeastern area in our region) should be handled to avoid a significant loss of relevant data. Future work will address these issues and aim to apply our method to larger regions for geologic interpretations based on the resulting linear database. We will also improve our method by developing new algorithms (e.g., to avoid the loss of adjacent data with similar orientations and lengths). Additionally, we hope to explore more advanced quantitative methods of evaluating similarity between linear datasets by using Hausdorff distances [20]. Once these improvements are made, generalization of our method for detecting patterns in other contexts will be another future research direction.

Acknowledgment

We would like to thank Meirman Syzdykbayev at the University of Pittsburgh for assisting with development of the initial Python code and Noah Villamarin-Cutter, Matt Gottfried, and Matt Cooney at Colorado College for their help working with the ArcPy library and debugging code.

References

[1] M. J. Smith and C. D. Clark, "Methods for the visualization of digital elevation models for landform mapping," *Earth Surf. Process. Landf.*, vol. 30, no. 7, pp. 885–900, Jul. 2005.

[2] M. J. Smith and S. M. Wise, "Problems of bias in mapping linear landforms from satellite imagery," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 9, no. 1, pp. 65–78, Feb. 2007.

[3] I. S. Evans, "Geomorphometry and landform mapping: What is a landform?," *Geospatial Technol. Geomorphol. Mapp. Proc. 41st Annu. Binghamt. Geomorphol. Symp.*, vol. 137, no. 1, pp. 94–106, Jan. 2012.

[4] J. Šilhavý, J. Minár, P. Mentlík, and J. Sládek, "A new artefacts resistant method for automatic lineament extraction using Multi-Hillshade Hierarchic Clustering (MHHC)," *Comput. Geosci.*, vol. 92, pp. 9–20, Jul. 2016.

[5] S. Solomon and W. Ghebreab, "Lineament characterization and their tectonic significance using Landsat TM data and field studies in the central highlands of Eritrea," *J. Afr. Earth Sci.*, vol. 46, no. 4, pp. 371–378, Nov. 2006.

[6] P. Štěpančíková, J. Stemberk, V. Vilímek, and B. Košťák, "Neotectonic development of drainage networks in the East Sudeten Mountains and monitoring of recent fault displacements (Czech Republic)," *Impact Act. Tecton. Uplift Fluv. Landsc. Drain. Dev.*, vol. 102, no. 1, pp. 68–80, Nov. 2008.

[7] A. Batayneh, H. Ghrefat, and A. Diabat, "Lineament characterization and their tectonic significance using gravity data and field studies in the Al-Jufr area, southeastern Jordan plateau," *J. Earth Sci.*, vol. 23, no. 6, pp. 873–880, 2012.

[8] D. Alegre Vaz, "Analysis of a Thaumasia Planum rift through automatic mapping and strain characterization of normal faults," *Geol. Mapp. Mars*, vol. 59, no. 11–12, pp. 1210–1221, Sep. 2011.

[9] U. Mallast, R. Gloaguen, S. Geyer, T. Rödiger, and C. Siebert, "Derivation of groundwater flow-paths based on semi-automatic extraction of lineaments from remote sensing data," *Hydrol Earth Syst Sci*, vol. 15, no. 8, pp. 2665–2678, Aug. 2011.

[10] D. Wladis, "Automatic lineament detection using digital elevation models with second derivative filters," *Photogramm. Eng. Remote Sens.*, vol. 65, pp. 453–458, 1999.

[11] A. A. Masoud and K. Koike, "Auto-detection and integration of tectonically significant lineaments from SRTM DEM and remotely-sensed geophysical data," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 6, pp. 818–832, Nov. 2011.

[12] G. Jordan and B. Schott, "Application of wavelet analysis to the study of spatial pattern of morphotectonic lineaments in digital terrain models. A case study," *Remote Sens. Environ.*, vol. 94, no. 1, pp. 31–38, Jan. 2005.

[13] D. U. Wise, R. Funiciello, M. Parotto, and F. Salvini, "Topographic lineament swarms: Clues to their origin from domain analysis of Italy," *Geol. Soc. Am. Bull.*, vol. 96, no. 7, pp. 952–967, Jul. 1985.

[14] B. Karimi, N. McQuarrie, J.-S. Lin, and W. Harbert, "Determining the geometry of the North Anatolian Fault East of the Marmara Sea through integrated stress modeling and remote sensing techniques," *Tectonophysics*, no. 0.

[15] Y. Kageyama and M. Nishida, "Lineament detection from land cover information in mixels using Landsat-TM data," *Electr. Eng. Jpn.*, vol. 148, no. 4, pp. 65–73, Sep. 2004.

[16] A. Abdullah, J. M. Akhir, and I. Abdullah, "Automatic mapping of lineaments using shaded relief images derived from digital elevation model (DEMs) in the Maran-Sungi Lembing area, Malaysia," *Electron. J. Geotech. Eng.*, vol. 15, pp. 1–9, 2010.

[17] P.C.I. Geomatics Enterprises, *Geomatica Help. [software help]*. Richmond Hill, Ontario, Canada, 2011.

[18] Environemental Systems Research Institute, *ArcMap*. Redlands, CA.

[19] H. Tveite, "An accuracy assessment method for geographical line data sets based on buffering," *Int. J. Geogr. Inf. Sci.*, vol. 13, no. 1, pp. 27–47, Jan. 1999.

[20] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Springer Science & Business Media, 2009.